Universidade de Aveiro Departamento de Eletrónica, Telecomunicações e Informática
2013

Mário Jorge Ferreira Rodrigues

Modelo de Acesso a Fontes em Linguagem Natural no Governo Eletrónico

Model of Access to Natural Language Sources in Electronic Government

Universidade de Aveiro
2013

Departamento de Eletrónica, Telecomunicações e Informática

Mário Jorge Ferreira Rodrigues

Modelo de Acesso a Fontes em Linguagem Natural no Governo Eletrónico

Model of Access to Natural Language Sources in Electronic Government

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Informática, realizada sob a orientação científica do Doutor António Joaquim Silva Teixeira, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro e do Doutor João Gonçalo Gomes de Paiva Dias, Professor Coordenador da Escola Superior de Tecnologia e Gestão de Águeda da Universidade de Aveiro.

Dedico este trabalho à Marlene, à minha mãe e ao meu pai que estão sempre comigo.

o júri

presidente

Doutor Armando da Costa Duarte
Professor Catedrático do Departamento de Química da Universidade de Aveiro

Doutor Joaquim Arnaldo Carvalho Martins
Professor Catedrático do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Doutor José Miguel de Oliveira Monteiro Sales Dias
Professor Associado Convidado do ISCTE-IUL – Instituto Universitário de Lisboa

Doutor Álvaro Manuel Reis da Rocha
Professor Associado Convidado do Instituto Universitário de Estudos e Desenvolvimento da Galiza da Universidade de Santiago de Compostela – Espanha

Doutor Alberto Manuel Brandão Simões
Professor Auxiliar Convidado do Instituto de Letras e Ciências Humanas da Universidade do Minho

Doutor António Joaquim da Silva Teixeira
Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro (Orientador)

Doutor João Gonçalo Gomes de Paiva Dias
Professor Coordenador da Escola Superior de Tecnologia e Gestão de Águeda da Universidade de Aveiro (Coorientador)

Doutora Liliana da Silva Ferreira
Senior Researcher, FhP-AICOS – Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions

agradecimentos

Aos meus orientadores, Professor António Teixeira e Professor Gonçalo Paiva Dias, pelo acompanhamento atento e pelo espírito de constante desafio

À Linguateca e à comunidade científica em geral pela disponibilização dos recursos e das ferramentas de software que permitiram implementar um protótipo funcional

Aos funcionários da Câmara Municipal de Águeda pelo tempo e atenção dispensados

Aos meus colegas de laboratório, em particular ao Aneesh Chauhan, pelas conversas acerca de tecnologia e de outros desafios deste percurso

Aos meus amigos e colegas da ESTGA, do IEETA e da UA em geral por terem contribuído com incentivos, sugestões e críticas e por terem colaborado nos testes para avaliação deste trabalho

À Marlene, minha mulher, e a toda a minha família pela sua preocupação, compreensão e afecto incondicional

palavras-chave

governo electrónico 2.0, informação semântica, ontologia, processamento de linguagem natural, extracção de informação baseada em ontologia, usabilidade de sistemas, modelo conceptual de fornecimento de informação para o governo electrónico.

Resumo

Para a efectiva existência de governo electrónico é necessário e crucial a disponibilização de informação e documentação pública e tornar simples o acesso a esta pelos cidadãos. Uma parte, não necessariamente pequena, destes documentos encontra-se sob uma forma não estruturada e em linguagem natural e, consequentemente, fora do que os sistemas de pesquisa actuais conseguem em geral suportar e disponibilizar eficazmente.

Assim, em tese, é possível melhorar o acesso a estes conteúdos com recurso a sistemas que processem linguagem natural e que sejam capazes de criar informação estruturada, em especial se suportados numa semântica. Com o objectivo de colocar esta tese à prova, o desenvolvimento deste trabalho integrou três grandes fases ou vertentes: (1) Criação de um modelo conceptual integrando a criação de informação estruturada e a sua disponibilização para vários actores, alinhado com a visão do governo electrónico 2.0; (2) Definição e desenvolvimento de um protótipo instanciando os módulos essenciais deste modelo conceptual, nomeadamente a extracção de informação suportada em ontologias e exemplos de informação relevante, gestão de conhecimento e acesso baseado em linguagem natural; (3) Uma avaliação de usabilidade e aceitabilidade da consulta à informação tornada possível pelo protótipo – e em consequência do modelo conceptual - por utilizadores num cenário realista e que incluiu comparação com formas de acesso existentes. Além desta avaliação, a outro nível, mais relacionado com avaliação de tecnologias e não do modelo, foram efectuadas avaliações do desempenho do subsistema responsável pela extracção de informação.

Os resultados da avaliação mostram que o modelo proposto foi percepcionado como mais eficaz e mais útil que as alternativas. Associado ao desempenho do protótipo a extrair informação dos documentos, comparável com o estado da arte, os resultados obtidos mostram a viabilidade e as vantagens, com a tecnologia actual, de utilizar processamento de linguagem natural e integração de informação semântica para melhorar acesso a conteúdos em linguagem natural e não estruturados.

O modelo conceptual e o protótipo demonstrador pretendem contribuir para a existência futura de sistemas de pesquisa mais sofisticados e adequados ao governo electrónico. Para existir transparência na governação, cidadania activa, maior agilidade na interacção com a administração pública, entre outros, é necessário que cidadãos e empresas tenham acesso rápido e fácil a informação oficial, mesmo que ela tenha sido originalmente criada em linguagem natural.

keywords

electronic government 2.0, semantic information, ontology, natural language processing,ontology based information extraction, system usability, conceptual model for information provision in electronic government.

abstract

For the actual existence of e-government it is necessary and crucial to provide public information and documentation, making its access simple to citizens. A portion, not necessarily small, of these documents is in an unstructured form and in natural language, and consequently outside of which the current search systems are generally able to cope and effectively handle.

Thus, in thesis, it is possible to improve access to these contents using systems that process natural language and create structured information, particularly if supported in semantics. In order to put this thesis to test, this work was developed in three major phases: (1) design of a conceptual model integrating the creation of structured information and making it available to various actors, in line with the vision of e-government 2.0; (2) definition and development of a prototype instantiating the key modules of this conceptual model, including ontology based information extraction supported by examples of relevant information, knowledge management and access based on natural language; (3) assessment of the usability and acceptability of querying information as made possible by the prototype - and in consequence of the conceptual model - by users in a realistic scenario, that included comparison with existing forms of access. In addition to this evaluation, at another level more related to technology assessment and not to the model, evaluations were made on the performance of the subsystem responsible for information extraction.

The evaluation results show that the proposed model was perceived as more effective and useful than the alternatives. Associated with the performance of the prototype to extract information from documents, comparable to the state of the art, results demonstrate the feasibility and advantages, with current technology, of using natural language processing and integration of semantic information to improve access to unstructured contents in natural language.

The conceptual model and the prototype demonstrator intend to contribute to the future existence of more sophisticated search systems that are also more suitable for e-government. To have transparency in governance, active citizenship, greater agility in the interaction with the public administration, among others, it is necessary that citizens and businesses have quick and easy access to official information, even if it was originally created in natural language.

Contents

1 Introduction **1**
- 1.1 Thesis Statement . . . 2
- 1.2 Research Objectives . . . 2
- 1.3 Contributions . . . 3
- 1.4 Publications . . . 4
- 1.5 Dissertation Structure . . . 6

2 Background and Related Work **9**
- 2.1 e-Government . . . 9
  - 2.1.1 Contributing Areas . . . 11
  - 2.1.2 Information and Knowledge Research in e-Government . . . 12
  - 2.1.3 Recent Initiatives . . . 14
- 2.2 Semantics and Human Language Technologies (HLT) . . . 15
  - 2.2.1 Knowledge Representation . . . 18
  - 2.2.2 Information Extraction . . . 19
  - 2.2.3 Natural Language Interfaces . . . 22
- 2.3 Semantics and HLT in e-Government . . . 24
- 2.4 Summary . . . 26

3 Conceptual Model **29**
- 3.1 A View of the Future of e-Government . . . 30
- 3.2 Requirements . . . 34
  - 3.2.1 Information Acquisition from Natural Language Documents . . . 34
  - 3.2.2 Data Storage using a Knowledge Base . . . 35
  - 3.2.3 Relevant Information Identification . . . 36
  - 3.2.4 Interoperability . . . 36
  - 3.2.5 User Interface . . . 37
- 3.3 Model Overview . . . 38
- 3.4 Usage Scenarios . . . 39
  - 3.4.1 Managed by an Information Producer . . . 41
  - 3.4.2 Managed by an Information Consumer . . . 43
  - 3.4.3 Managed by a Third Party . . . 44
- 3.5 Summary . . . 44

4 Resources and Tools **47**
- 4.1 Resources . . . 47
  - 4.1.1 Annotated Corpus . . . 48
  - 4.1.2 Computational Lexicon . . . 49

4.1.3 Lexical Relations Database . . . 50
4.1.4 Ontologies . . . 50
4.2 Tools . . . 53
4.2.1 Sentence Boundary Detection . . . 54
4.2.2 Part-of-Speech Tagging . . . 56
4.2.3 Named Entity Recognition . . . 60
4.2.4 Syntactic Parsing . . . 63
4.2.5 Ontology Creation and Edition . . . 65
4.2.6 Semantic Annotation . . . 68
4.2.7 Natural Language Interface . . . 70
4.3 Summary . . . 72

5 Prototype Implementation **73**
5.1 Prototype Overview . . . 73
5.2 Natural Language Processing . . . 74
5.3 Domain Representation . . . 78
5.4 Semantic Extraction and Integration . . . 80
5.5 Summary . . . 84

6 Evaluation Results **87**
6.1 Usability Assessment . . . 88
6.1.1 Methodology . . . 91
6.1.2 Experiment Setup and Participants . . . 93
6.1.3 Results . . . 94
6.1.4 Discussion . . . 101
6.2 Information Extraction Performance . . . 103
6.2.1 Methodology . . . 104
6.2.2 Results . . . 105
6.2.3 Discussion . . . 106
6.3 Application Examples . . . 106
6.3.1 Internal Organization Example . . . 106
6.3.2 Transparency Example . . . 107
6.3.3 Citizen Example . . . 108
6.3.4 Other Concerns such as Activity Indicator . . . 109
6.4 Summary . . . 111

7 Conclusions **113**
7.1 Work Overview . . . 113
7.1.1 Early Study and Exploratory Tests . . . 113
7.1.2 Conceptual Model Proposal and Prototype Development . . . 115
7.1.3 Usability Tests with External Subjects . . . 116
7.2 Discussion . . . 117
7.2.1 Strengths of the Approach . . . 118
7.2.2 Weaknesses of the Approach . . . 119
7.3 Future Directions . . . 120
7.4 Epilogue . . . 121

A LABEL-LEX-sw Tag Set Conversion **143**

B Semantic Model Creation Algorithm **145**

C Application Setup **147**

D Product Reaction Cards Translation **149**

List of Figures

2.1 Research areas contributing to the e-government field of applied research. . . . 12
2.2 Difference between information extraction and information retrieval. 20
2.3 Possible syntactic structures of a sentence. 21

3.1 Political scenarios outlined by Frissen et al. (2007). 31
3.2 UNPAN four stage model to measure e-government evolution. 33
3.3 Proposed conceptual model. 40
3.4 Examples of service provision sophistication at three government levels. 41
3.5 System managed by the information producer scenario. 42
3.6 System managed by the information consumer scenario. 43
3.7 System managed by a third party scenario. 44

4.1 Example of a possible TreeTagger decision tree. 58
4.2 Example of a sentence dependency structure 63
4.3 Screenshot presenting the ontology visualizer. 66
4.4 Screenshot presenting a semantic annotation interface. 70

5.1 Proof-of-concept prototype architecture. 74
5.2 Sentence graph corresponding to a NLP example output. 77
5.3 Graph of the syntactic structure of the fragment presented in Table 5.1. 85

6.1 Screenshot of the prototype natural language interface. 92
6.2 Screenshot of Loures municipality website. 93
6.3 Screenshot of Sintra municipality website. 94
6.4 Number of cards selected by participants. 96
6.5 Proportion of positive and neutral cards selected. 99
6.6 Frequency of cards selected as top 5 for the prototype. 100
6.7 Cards selected as top 5 of the prototype. 101
6.8 Frequency of cards selected as top 5 of search systems of municipalities. 102
6.9 Cards selected as top 5 of search systems of municipalities. 102
6.10 Screenshot of a website page rendering a query output in a map. 109
6.11 Map rendering part of Arouca municipality for which information was found. . 110

List of Tables

2.1 Relevant related research projects about government information provision. . . 16
2.2 Some currently deployed initiatives lead by public sectors. 17
2.3 Comparison of ontology based information extraction systems. 23
2.4 Semantic interoperability standard definition efforts lead by governments. . . . 27

3.1 Characterization of the political scenarios outlined by Frissen et al. (2007). . . . 30

4.1 Description of CoNLL-X shared task token fields. 48
4.2 First 20 lines of Bosque v7.3 in CoNLL-X format. 49
4.3 First 10 lines of LABEL-LEX-sw. 50
4.4 Sample of PAPEL relation set. 51
4.5 First 10 lines of the file with all relations of PAPELv3.2. 51
4.6 The 15 elements of Dublin Core Metadata Initiative. 53
4.7 Geo-Net-PT features. 53
4.8 Summary of available sentence boundary detector software tools 55
4.9 Precision achieved by state-of-the-art POS taggers with different corpora. . . . 57
4.10 Samples of files used to train TreeTagger. 59
4.11 Tag set used to train TreeTagger. 59
4.12 Summary of named entity recognition systems working for Portuguese. 61
4.13 Classes and types used in the second HAREM. 62
4.14 Results of the CoNLL-X shared task. 64
4.15 Properties of some ontology editors listed in W3C website. 67
4.16 Relevant semantic annotators. 69
4.17 Comparison of state-of-the-art natural language interfaces. 71

5.1 Example of the output of the system modules. 75
5.2 CPOSTAG field value assigned in function of TreeTagger output POS tag. . . . 77
5.3 Subclasses of the class `ExecutiveSubject`. 79
5.4 Relations which domain is exclusively `ExecutiveSubject`. 79
5.5 Example of a pre-annotation configuration file. 80

6.1 The complete set of product reaction cards. 89
6.2 Top ten most populated Portuguese municipalities, Portuguese census 2011. . . 95
6.3 Sentiment distribution of all cards selected. 97
6.4 List of cards not selected by participants. 98
6.5 Sentiment distribution of participants' top 5 choices. 100
6.6 Information detected by the system. 106
6.7 SPARQL query relative to protocols and respective explanation. 107
6.8 Number of protocols between a municipality and other institutions. 108

6.9 SPARQL query about which persons applied a building permit. 110
6.10 Status of the building permits requested by citizens which name includes *Maria*. 110

A.1 Conversion of LABEL-LEX-sw tags to CoNLL-X (Bosque) tags. 144

D.1 Translation of product reaction cards. 149

Glossary

API Application Programming Interface

CNL Controlled Natural Language

CoNLL-X Tenth Conference on Computational Natural Language Learning

DAML DARPA Agent Markup Language

DCMI Dublin Core Metadata Initiative

DR Domain Representation

EU European Union

FOAF Friend of a Friend

GUI Graphical User Interface

HLT Human Language Technologies

HTML HyperText Markup Language

ICT Information and Communication Technologies

IE Information Extraction

IEEE Institute of Electrical and Electronics Engineers

GPS Global Positioning System

KB Knowledge Base

k-NN k-Nearest Neighbor

KR Knowledge Representation

LAS Labeled Attachment Score

NER Named Entity Recognition

NLI Natural Language Interface

NLP Natural Language Processing

NUTS Nomenclature of Territorial Units for Statistics

OBIE Ontology-Based Information Extraction

OECD Organisation for Economic Co-operation and Development

OIL Ontology Inference Layer

OWL Web Ontology Language

PAPEL Palavras Associadas Porto Editora Linguateca

PDF Portable Document Format

POS Part-Of-Speech

RDF Resource Description Framework

RDFS Resource Description Framework Schema

SEI Semantic Extraction and Integration

SHOE Simple HTML Ontology Extensions

SPARQL Simple Protocol and RDF Query Language

SQL Structured Query Language

SKOS Simple Knowledge Organization System

SWRL Semantic Web Rule Language

UAS Unlabeled Attachement Score

W3C World Wide Web Consortium

WGS84 World Geodetic System 1984

WSJ Wall Street Journal

WWW World Wide Web

XML eXtensible Markup Language

1 Introduction

Governments are knowledge intensive organizations and are expected to become increasingly so in the future (Osimo, 2008). Among the member countries of Organisation for Economic Co-operation and Development (OECD), the role of governments in society has increased over the last decades (OECD, 2005). The diffusion of Information and Communication Technologies (ICT) has fostered economic growth and social progress in the past few decades, as well as redefined how citizens and businesses relate to each other and to government (Ho et al., 2011; Seo et al., 2009; Heeks, 2008). Society is changing its attitude and expectation towards a more open government. The expectation has evolved from hope to demand, and even to legal right to access information. This contributes to empower citizens and businesses which is seen as the trend to have the most significant impact in the coming decades (Frissen et al., 2007).

e-Government is a term coined to refer to the use of ICT by government. The view of e-government 2.0 – e-government platforms that take advantage of Web 2.0 technologies – is about having government as a platform, a provider of data and services. For this to happen it is necessary that governments provide data in a non-proprietary and predictable formats. Content has to be addressable in a granular form, in formats that are open, structured and machine-readable (United Nations, 2010; Chang and Kannan, 2008). However, a relevant set of government related documentation is originally created in natural language plain text documents. Natural language is a convenient and adequate form of making information available – government has to serve all population – but it is also a challenge for automatic computerized access to contents as natural language is extensive, ambiguous and contains personal

differences in vocabulary and style.

Human Language Technologies (HLT) refers to speech technology and Natural Language Processing (NLP). NLP was defined by Allen (2000) as "computer systems that analyze, attempt to understand, or produce one or more human languages, such as English, Japanese, Italian, or Russian". Understanding text implies to have a semantically informed interpretation and abstraction of the content. This makes possible to relate information, combine different sources, and provide more accurate query tools Guarino (1998). When applied in e-government context it can bring clear benefits for citizens, businesses and other branches of government. However, the use of HLT has not been methodical or frequent in the e-government context.

1.1 Thesis Statement

The use of semantic and NLP technologies can improve government information access, and thus increase society empowerment, without compromising government's important concerns - serve 100% of population, reduce or eliminate digital divide and information asymmetry - or requiring a significant change in the way government services operate and produce information.

1.2 Research Objectives

The main goal of this research is to evaluate if the use of semantic and NLP technologies contributes to improve the access to information contained in natural language government documents. To achieve this broad goal, four main objectives were defined:

Identify the needs and restrictions of information provision in the e-government environment. This research objective is a necessary condition when designing and developing systems able work in any environment and also in e-government. Despite the little amount of data available about the success rates of early e-government initiatives, which is imputed to political implications, some authors advance a baseline estimation around 70%-80% of total or partial failures (Misuraca, 2009). They also advance that the main cause of failures seems to be the lack of fully understanding the complex nature of e-government and the exact extent of its challenges and constraints. This is an important issue as it can make a difference in the adoption of new solutions.

Propose a conceptual model able to provide information of natural language documents and suitable for e-government. This objective is a proposal for a generic architecture observing the findings of the first objective. The architecture should be generic as it aims to be usable in a wide range of government subjects and to be adaptable to a wide range of natural languages. This research objective is of crucial importance since, as Leigh-

ninger (2011) observed, the real value to be added in e-government platforms is not the use of temporary tools but the building of a more durable infrastructure for participatory democracy.

Develop a proof-of-concept prototype instantiating the conceptual model and working for Portuguese documents. Developing of a proof-of-concept prototype has impact from three perspectives. The first is to verify and fine tune the proposed conceptual model. After a top-down model proposal emphasizing a complete understanding of the system follows a fine tuning step using a bottom-up approach to make the model architecture aligned with existing software tools. The fine tuning aims to make easier for others to implement systems complying with the proposed model architecture. The second perspective is to provide an implementation guideline and also allow testing if the prototype improves the access to information contained in natural language government documents. The third reason is to contribute with a working application for the research community. The Portuguese research community has been very active in producing tools, namely for NLP, but there is a lack of applications to use and improve upon.

Evaluate the approach using a realistic scenario. After studying the needs and restrictions of e-government, and proposing a conceptual architecture, it is important to define the context where the architecture is valid. This is done by defining usage scenarios. Testing the prototype behavior in a defined scenario, with real documents and subjects not involved in the project, configures a realistic scenario helpful to evaluate the approach, detect early prototype failures and propose improvements. It is important to take users' feelings into account in the system design. If users do not experience satisfaction, their attitude about using the system may affect the efficiency and effectiveness of their interaction with it (Barnum and Palmer, 2011). The success of e-government initiatives may be dictated by the amount of effort people are obliged to make when using its systems.

The third and fourth objectives correspond to the thesis statement part: "*The use of semantic and NLP technologies can improve government information access, and thus increase society empowerment...*" while the first two objectives correspond to: "*... without compromising government's important concerns ... nor requiring a significant change in the way government services operate and produce information*".

1.3 Contributions

This thesis presents a study and a novel prototype for automatic acquisition and structuring of information from natural language government documents written in Portuguese. The prototype makes that information available in formats suitable for access by people and

machines. Its architecture follows a conceptual model designed taking into consideration e-government restrictions and needs. Besides the contributions resultant from the fulfillment of all research objectives, the development of the prototype brought some other innovative contributions to the fields of e-government and Information Extraction (IE) in Portuguese texts, more specifically Ontology-Based Information Extraction (OBIE), and complex system evaluation. These can be summarized as follows:

1. A conceptual model for natural language information acquisition and provision suitable for the e-government context. In the process of designing the conceptual model some self-contained contributions were made:
   (a) The outline and explanation of three possible usage scenarios for e-government information systems.
   (b) A detailed discussion on why and how the use of HLT can contribute to solve some generic challenges of e-government. This discussion was published in the form of a position paper (Rodrigues et al., 2010b).
   (c) A detailed discussion on why and how e-government information systems are important and can have a positive impact on businesses, namely through enterprise 2.0 platforms. This discussion was published as a book chapter (Rodrigues et al., 2013).
2. A new OBIE working prototype for Portuguese language. From the development of the prototype some resources were created and are ready to be available to the community:
   (a) A Portuguese ontology for municipal subjects. This ontology is compatible with a natural language interface allowing information search using Portuguese natural language (Kaufmann et al., 2007).
   (b) A Part-Of-Speech (POS) tagger model for Portuguese. The model is n-gram ready to be used with TreeTagger (Schmid, 1994).
   (c) A dependency grammar for syntactic parsing of Portuguese. The grammar is ready to be used with MaltParser (Hall et al., 2007).
3. The creation of tools in Portuguese to evaluate the usability and acceptability of complex systems. These tools were already used to evaluate other projects and are based on Microsoft's Product Reaction Cards (Benedek and Miner, 2002).

1.4 Publications

Along the study and development of the conceptual model and working prototype, several publications were made about the research findings and performance results of the proof-of-concept prototype. In chronological order:

- A position paper was published arguing that the inclusion of HLT in e-government is still a challenge but its benefits are clear: reduction of the digital divide; more intuitive human interfaces; communication comparable to traditional face-to-face dialogs; more channels available; and more knowledgeable systems. These benefits are particularly relevant in e-government because government must serve all the population and because e-government success also depends on how easy it is to use its systems.

  Rodrigues, M., Dias, G. P., Teixeira, A., April 2010b. Human language technologies for e-gov. In: Filipe, J., Cordeiro, J. (Eds.), WEBIST 2010, Proceedings of the 6th International Conference on Web Information Systems and Technologies. Organized by INSTICC in cooperation with WfMC and ACM SIGMIS, Valencia, Spain, pp. 400–403.

- A paper was published reporting the results of an early experiment on the process of geocoding entities referred in natural language texts. The approach detected entities that have some kind of address clues in their name or that exist in the ontology GeoNetPT-01 (Chaves et al., 2005). The acquired information was stored in a relational database and the access to information is through a web page displaying a world map with marks on top of the locations referred in texts.

  Rodrigues, M., Dias, G. P., Teixeira, A., June 2010a. Automatic extraction and representation of geographic entities in egovernment. In: Rocha, A., Sexto, C. F., Reis, L. P., Cota, M. P. (Eds.), Sistemas y Tecnologías de Informacíon - Actas de la 5a Conferencia Ibérica de Sistemas y Tecnologías de Informacíon. AISTI, GIS-T and USC, Santiago de Compostela, Spain, pp. 160–163.

- The third publication was a paper outlining the first version of the architecture of the IE prototype. At the time, the working prototype acquired information about a single topic, municipal subsidies, which was integrated with geographical information. All information was stored in a knowledge base conforming to an ontology. This version of the system used a fixed set of rules to extract information, and thus a change in the ontology most certainly would imply a redefinition of those rules.

  Rodrigues, M., Dias, G. P., Teixeira, A., November 2010c. Knowledge extraction from minutes of portuguese municipalities meetings. In: Mateo, C. G., Diaz, F. C., Pazó, F. M. (Eds.), FALA 2010: VI Jornadas en Tecnologia del Habla and II Iberial SLTech - Speech and Language Technologies for Iberian Languages. MTG, RTTH and ISCA SIG-IL, Vigo, Spain, pp. 51–54.

- Another paper reported an improved prototype version, now using an OBIE approach, which was a generalization of the previous one. The generalization included the abil-

ity to extract generic, user selected information without any software reconfiguration. The information extraction was learned from examples of associations between ontology classes and relations and natural language sentences, defined using a Graphical User Interface (GUI). The ontology was defined in Web Ontology Language (OWL) and, when changed, it was necessary to give new examples of associations between ontology classes and relations and natural language sentences. There was no need to reprogram any rules whatsoever.

Rodrigues, M., Dias, G. P., Teixeira, A., October 2011b. Ontology Driven Knowledge Extraction System with Application in e-Government. In: Proc. of the 15th Portuguese Conference on Artificial Intelligence. Lisboa, Portugal, pp. 760–774.

- A journal article was published adding, to the OBIE module, the description of the newly added interfaces: one for people to access information via web and via natural language, and another for third party systems query the knowledge base using Simple Protocol and RDF Query Language (SPARQL).

  Rodrigues, M., Dias, G. P., Teixeira, A., 2011a. Criação e Acesso a Informação Semântica Aplicada ao Governo Eletrónico. Linguamática 3 (2), 55–68.

- A book chapter discusses the impacts of government in some challenges faced by Enterprise 2.0. It provides an overview of e-government current restrictions and guidelines to future directions. In this book chapter it is argued that a conceptual model architecture as the one proposed here answers to current e-government concerns and can bring important benefits for Enterprise 2.0 systems.

  Rodrigues, M., Dias, G. P., Teixeira, A., 2013. Towards e-government information platforms for enterprise 2.0. In: Cruz-Cunha, M. M., Moreira, F., ao Varajão, J. (Eds.), Handbook of Research on Enterprise 2.0: Technological, Social, and Organizational Dimension. IGI Global.

1.5 Dissertation Structure

This dissertation is organized in seven chapters. It starts with the present chapter which introduces the research objectives, the thesis statement and contributions, results already published, and the dissertation structure. After follow chapters 2 to 7:

- The second chapter introduces relevant background information and related work. It starts by summarizing e-government opportunities and challenges, and includes sections about Knowledge Representation (KR), IE, and OBIE.

- The third chapter presents a conceptual model for extracting structured semantic information from unstructured, natural language (government) texts. It provides a view to the future of e-government, the scenarios that were outlined, the requirements of the system and the general software architecture.
- The fourth chapter describes the resources and software tools used in the prototype implementation.
- The fifth chapter presents the implemented prototype and describes how the software tools were configured and integrated in a coherent system. It contains sections that elaborate on each of the prototype modules: NLP, KR, and Semantic Extraction and Integration (SEI).
- The sixth chapter discusses the methods used to measure the system performance and provides some example applications. The chapter begins by reporting a system usability test to find how potential users feel when using the prototype. After is described the performance evaluation in terms of precision and recall - two common measures in the NLP area - and continues with example applications that benchmark the impact of the system against current solutions.
- The seventh chapter concludes the document with considerations about the work contributions, its achievements and its limitations. The chapter ends with an outline for future work.

2

Background and Related Work

This chapter covers the background areas and related work necessary to understand the contributions of this dissertation. It discusses the current state in the fields of e-government, semantic representation, Human Language Technologies (HLT), and the application of semantic and HLT to the e-government domain. It starts by introducing the area of e-government with an explanation of its different perspectives and an overview of some recent initiatives relative to government information provision and standard definition. Then it is provided a broad description of HLT followed by a discussion on the topics relevant for this work, specifically: Knowledge Representation (KR), Information Extraction (IE), Ontology-Based Information Extraction (OBIE), and Natural Language Interface (NLI). The chapter ends discussing the application of HLT to the e-government domain.

2.1 e-Government

e-Government is a dynamic concept of varying meaning and significance (Relyea, 2002). Without a universally accepted definition, e-government broadly refers to the use of ICT by government agencies to improve their interactions with citizens, businesses, and other branches of government. The use of ICT in government contributed to make most administrations more efficient, more flexible, more transparent and oriented to its customers: citizens, businesses, other branches of government (European Commission, 2011; United Nations, 2010; OECD, 2005). The definition of e-government provided by World Bank (2011) is comprehensive and shares the essence of many other (for instance see Jeff (2000); Relyea (2002); OECD (2005);

United Nations (2010)):

"*E-Government refers to the use by government agencies of information technologies (such as Wide Area Networks, the Internet, and mobile computing) that have the ability to transform relations with citizens, businesses, and other arms of government. These technologies can serve a variety of different ends: better delivery of government services to citizens, improved interactions with business and industry, citizen empowerment through access to information, or more efficient government management. The resulting benefits can be less corruption, increased transparency, greater convenience, revenue growth, and/or cost reductions. ...*"

The use of technology in government has been studied for more than thirty years. An article by Ayres and Kettinger (1983) discusses the improvement of government workers' productivity through the use of computing and networking technologies. However, the concept associated to e-government, that technology can be used to improve interactions between government and society, appears to be introduced in 1990 decade. When discussing the existence of a public interest in the United States National Information Infrastructure, Weingarten (1994) identifies three categories: (1) public applications, that includes among others the "dissemination of government information and delivery of government services"; (2) equity of access, to reduce the gap between the "haves" and the unempowered trough the "access to resources and services over the Net", among other strategies; (3) information rights, "protecting basic rights such as freedom of speech, privacy, intellectual property, and access to public information".

Other authors followed discussing the potential policies and practices needed for the "electronic government of the future" (see for example Belanger and Carter (2012) for an historical review on e-government research in information systems). Over time, the research questions and practices of e-government have been sensitive to factors like the plurality of political views, political will, societal and economical changes, and technological progress (Hendriks, 2003; Hardy and Williams, 2011). Today, e-government is reckoned as an important research area by researchers and practitioners alike and, consequently, several journals and conferences are dedicated to this topic. The importance of this area derives from government's influence over the society through legislation, regulation, taxation, definition of the conditions under which firms compete, and also as a client and a service provider. Adding to this widespread reach, is the influence of local governments (Chang and Kannan, 2008; Carroll and Buchholtz, 2011).

However, the deployment of e-government initiatives to the general public has not been as successful as expected. Some authors estimated that early e-government initiatives had a percentage of failed projects around 70% to 85% (Fountain, 2001; Heeks, 2003; Collins, 2007). These estimates are rough since there is little amount of data to arrive to these numbers. This scarcity of data is imputed to the political implications of such numbers (Misuraca,

2009). The main causes of failure are attributed to the lack of preparation, e-readiness, and poor assessment of the real needs causing unrealistic assumptions. Adding to these, there is a problem in fully understanding the exact extent of government information systems challenges and constraints. Governments are complex organizations with bureaucratic systems designed to have checks and balances guaranteeing that powers are divided into separate levels and branches of government (Scholl and Klischewski, 2007).

2.1.1 Contributing Areas

e-Government is a broad and interdisciplinary field of applied research that gathers contributions from several areas. While several authors refer to this, Codagnone and Wimmer (2007) elaborated further and observed that the contributions mainly come from five research areas (see Figure 2.1):

Political, democracy, and legal sciences - studies the impacts of ICT usage in the course of decision-making at political, strategic and governance level. Some dimensions investigated by these disciplines are ICT supported democracy, direct participation of citizens, elected representatives support in democratic decision making, co-governance between state institutions and civic actors, and community involvement. Recent topics are the study of how social networks can be used to increase communication between citizens and elected officials.

Organizational, public administration, and economic sciences - researches how governance structures can be organized to increase efficiency and effectiveness of the public sector, and the effect of such organizations on productivity, efficiency and compliance to laws. Proposed studies and solutions include networked governments, public-public, public-private, and public-private-civic partnerships. It aims to increase governments' quality of service and public value generation by making them more accountable and transparent. Monitoring and benchmarking are key methodologies for this domain of research.

Computer sciences - focus on the concepts and solutions for ICT implementation. Prominent examples of research aspects are interoperability between bureaucratic systems at regional, national and international levels; multi-channel public service provision; electronic identification, encryption, digital signatures, and electronic payments. Two important and related concerns in this area are the definition of standards and avoiding vendor lock-in.

Information and knowledge research sciences - studies how information and knowledge resources should be managed taking into consideration concerns about transparency, trust, privacy, compliance with the laws, and accessibility. The goal is to provide intelligent and ubiquitous support systems for decision making, services provision, employee

knowledge portals, policy discussions, citizen participation, etc. Research topics include intelligent content, knowledge sharing mechanisms and comprehensive one-stop accessibility of dispersed knowledge resources.

Social and human sciences - contributes to a deep and articulated definition of citizens, businesses, and other branches of government. It studies the variety of stakeholders and institutional aspects including its motivations, how citizens and businesses interact with governments, how governments can establish better and more trust based relations with their constituencies, and how employees interact within and across organizations. This area also studies the impacts of using ICT in government in terms of social exclusion and/or inclusion and in terms of economic value creation.

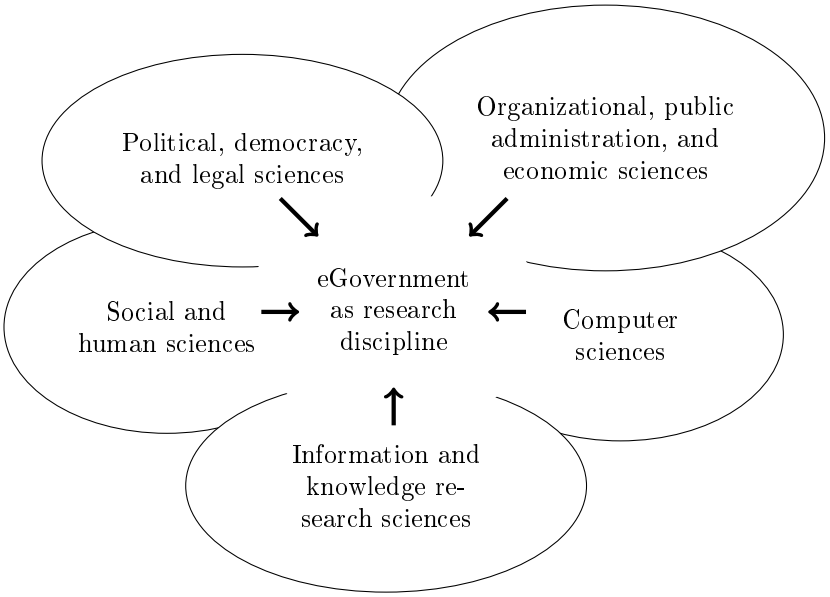


Figure 2.1: Research areas contributing to the e-government field of applied research. Source Codagnone and Wimmer (2007, p. 14).

Having contributions from several research areas implies that to research in e-government is necessary to understand the interplay of the many aspects in a holistic way as a socio-technical system (Codagnone and Wimmer, 2007). We will now focus the discussion on the e-government contributing research area most relevant for this work: information and knowledge research sciences.

2.1.2 Information and Knowledge Research in e-Government

Among other, e-government has the potential to contribute to citizen empowerment, and government transparency and accountability by providing tools to make public information - laws, regulations, service instructions, deliberations, etc. - easily accessible to every citizen

and business, anytime, anywhere and using whatever platform. Dias and Moreira (2008) assert that "Public access to relevant information, including the studies that sustain and the rules that superintend both bureaucratic and political decisions, help to promote a more informed, more aware and less condescending, citizenship". Disclosing information about government initiatives, what was made and why was made, the achieved or expected results, and how citizens and businesses can benefit from them, improves government transparency, accountability, effectiveness and efficiency. The absence of information fuels people's perceptions that services are wasteful and inefficient, and slowly erodes their trust in government and their commitment to citizenship (McQueen et al., 2009).

The diffusion of ICT has fostered economic growth and social progress in the past few decades, as well as redefined how citizens and businesses relate to each other and to government (Ho et al., 2011; Seo et al., 2009; Heeks, 2008). Citizens and businesses are changing their attitude and expectation towards a more open government. In recent years, society has increasingly demanded to know what decisions have been taken and by who. The expectation is that citizens and businesses are made aware in advance and consulted about decisions that affect them. This expectation has evolved from hope to demand, and even to legal right for access to information. Nowadays, there is a legal right to some types of information in most OECD member countries. In the past, governments chose what to reveal, now, the principle is becoming to make available all government data unless there is a defined public interest in withheld it (OECD, 2005). This brings empowerment to citizens and businesses and is seen as the trend to have the most significant impact in the coming decades (Frissen et al., 2007). Surveys across Europe (Colclough and Tinholt, 2009), United States (Reddick, 2005) and other countries (Andersen et al., 2010; Dias, 2011) show that it's still necessary to improve the capacity of providing government's relevant information to society.

Osimo (2008) acknowledges that governments are knowledge-intensive organizations, that they will become increasingly so in the future and that knowledge management is key to improving the efficiency and effectiveness of government. However, sharing data online is not sufficient to effectively share knowledge. It is necessary to organize the information, provide a semantic description and use open standards that will help information consumers to absorb, analyze, and interpret it. As an effect, governments are trying to take advantage of the success of Web 2.0 technologies (blogs, wikis, RSS feeds, social software, folksonomies, etc.) as a way to avoid mistakes of earlier and unsuccessful initiatives and understand the barriers to e-government (Misuraca, 2009). This trend is named e-government 2.0.

Alongside an optimistic view related to the use of web 2.0 technologies in e-government, the research community reckons that there are some risks. These risks assume particular relevance in the government context because of its institutional role and universal service obligations (Osimo, 2008; Hogben, 2007). Among them are:

1. Low participation: the usage of web 2.0 tools does not lead automatically to greater

user involvement;

2. Participation restricted to elite: most web 2.0 applications are used by the cultural and economic elite and can make societal divides wider by giving more voice to those that already have it. Digital divide refers to the gap between people with effective access and capacity to use ICT and those with very limited or no access or capacity at all;

3. Loss of control due to excessive transparency: in a few cases, opening-up the conversation has led to loss of control and loss of credibility. Cases were reported where sensitive information was released in an incorrect and sometimes illegal way;

4. Manipulation of content by interested parties: there is a concern that when social media become mainstream, vested interests would take over the content production.

5. Privacy and security issues: web 2.0 users appear not to be fully aware of the implications of publishing their details on the web, and web 2.0 applications in the government context could become a further source of sensitive information being published.

Another related concern, named information asymmetry, is raised by Clarkson et al. (2007). Information asymmetry is the lack of information awareness. It is not about the possession of means and knowledge to access information but is about knowing that the information exists. Clarkson et al. (2007) explains that "information asymmetry exists when a party or parties possess greater informational awareness pertinent to effective participation in a given situation relative to other participating parties". This problem can be reduced by making systematic and increased information sharing in order to correctly regulate all participants' expectations.

The above issues affect all tiers of government: national, regional, and local. Adding to those, evidence gathered in the 2010 European e-government benchmark survey show that local and regional e-government services are lagging behind when compared to central government (Lörincz et al., 2010). Corroborating these evidences, Santos and Amaral (2008) as well as Dias (2011) conducted studies for municipalities of Portugal, a country that has good classifications in international e-government benchmarks at national level services. These studies found that despite the investment made in the last decades in e-government, Portuguese municipalities still exhibit medium level development in what relates information dissemination through the web. This issue is particularly relevant as municipalities are often the closest point of service for citizens and enterprises (Rodrigues et al., 2010b).

2.1.3 Recent Initiatives

Osimo (2008) presents a study on e-government that addresses four general questions: (1) Are web 2.0 applications relevant for the government context? (2) If they are, in what way is web 2.0 likely to have an impact on government? (3) How significant could this impact be? (4) How are web 2.0 applications implemented in the government context?

The study concludes that there is sufficient evidence that web 2.0 applications are relevant for many different domains of e-government and those web 2.0 applications contribute to the key goals of better, simpler, joined-up and networked government. Also, the wide availability of public data for re-use seems to be an important enabling factor for web 2.0 applications and the managers of these initiatives agreed that wider availability of public data was their main recommendation to policy makers.

As the use of web 2.0 technologies in e-government, or e-government 2.0, is likely the trend of the near future, here are presented some relevant e-government information and knowledge researches that take advantage of web 2.0 technologies. These initiatives have been promoted by the academic community as well as by practitioners. The research community proposals range from conceptual works, like framework proposals, to more practical works such as web portals and proof-of-concept prototypes. Table 2.1 summarizes some relevant research projects about government information provision.

The works developed by practitioners are typically web portals that collect, merge, and make available government data. A summary of some relevant initiatives lead by public sectors is presented in Table 2.2. One of the initiatives presented is the `Data.gov` website, an official website of the United States Government, which serves as a single point of access to public data. Launched in 2009, it is one of the most substantial steps taken to provide a data platform for third parties. It functions as a clearinghouse for datasets generated by the government in an accessible and developer friendly format. Similar steps were later taken in other places as the United Kingdom and the State of São Paulo in Brazil (McQueen et al., 2009).

2.2 Semantics and Human Language Technologies (HLT)

Semantics is the study of meaning of linguistic expressions (Lewis, 1970). It studies the relations between signifiers, such as words, phrases, signs and symbols, and what they stand for. The language can be a natural language, such as English or Portuguese, or an artificial language, like a computer programming language or mathematics.

Human Language Technologies (HLT) refer to speech and Natural Language Processing (NLP). Research and development in speech processing includes automatic speech recognition to get a textual representation of a speech sound wave, and text to speech to generate a speech sound wave representing a given text. Allen (2000) defines NLP as referring "to computer systems that analyze, attempt to understand, or produce one or more human languages, such as English, Japanese, Italian, or Russian. The input might be text, spoken language, or keyboard input. The task might be to translate to another language, to comprehend and represent the content of text, to build a database or generate summaries, or to maintain a dialogue with a user as part of an interface for database/information retrieval". HLT can be divided in several subareas, such as Computational Linguistics, Information Extraction, Infor-

Table 2.1: Relevant related research projects about government information provision.

Reference	Challenge	Approach	Outcomes
Chen (2012)	Incorporates a wide variety of factors from previous research into a single theoretical framework.	A framework for Government 2.0 development and implementation.	No practical implementation was discussed.
Hobson et al. (2011)	Municipal employees rely heavily on manual methods for data sharing.	Improve interaction and information sharing within and between municipal departments.	No practical implementation was discussed.
Wang et al. (2011)	Support ecological and environmental issues such as biodiversity loss, water problems.	Monitoring systems based on linked data approach.	Web portal.
Fonou-Dombeu and Huisman (2011)	Strengthen the adoption of semantic technologies in e-government.	Combine ontology building methodology with state-of-the-art semantic web platforms.	Government domain ontology.
Mohamed et al. (2011)	Improve conventional development processes in terms of time efficiency and collaboration.	Conceptual software development technique for building and developing e-government 2.0 portal.	Proof-of-concept prototype.
Goodwin et al. (2008)	Represent links between geographic entities in spatial queries.	Encoding topological relations between geographic entities over traditional spatial queries.	Example dataset for the administrative geography.
Wang et al. (2007b)	Multidimensional information integration, especially spatial information integration.	Centralized database approach, accepting various structured data sources and data formats.	Integrated multidimensional centralized database.

Table 2.2: Some currently deployed initiatives lead by public sectors.

Name	Domains of usage in government	Address
Apps for Democracy USA	Makes government data useful for citizens, businesses and government agencies.	http://www.appsfordemocracy.org
Data.gov USA	Improves access to the data generated by the government, uses Linked Data.	http://www.data.gov
Eurostat EU	Improves access to the data generated by the government, datasets organized by themes but not semantically described.	http://epp.eurostat.ec.europa.eu
OpenPSI UK	Supports government based information publishers, research communities, and web developers.	http://www.jisc.ac.uk
PEPPOL EU	Enables businesses to easily deal electronically with any European public sector buyers in their procurement processes.	http://www.peppol.eu
Regulations.gov USA	A source for U.S. government regulations and related documents.	http://www.regulations.gov
USASpending USA	Publishes data about government spending, increasing data quality by adopting a uniform method of identifying the recipients of federal funds.	http://www.usaspending.gov
– Canada	Provides driving conditions and advice to drivers and commuters, combines Mapquest data with realtime traffic data to create a "mashup".	http://www.gov.bc.ca/tran

mation Retrieval, Language Understanding and Language Generation (Jurafsky and Martin, 2008).

Semantics and HLT overlap when studying the meaning of natural languages. Without pretending to provide a complete view of these two broad and rich research fields, the next subsections focus on some relevant areas for this work, areas where these two fields connect and overlap: Knowledge Representation (KR), Ontology-Based Information Extraction (OBIE) and Natural Language Interface (NLI). As both OBIE and NLI use ontologies at their core, the first subsection introduces the problem of Knowledge Representation (KR) and ontologies. The second subsection starts by the general problem of Information Extraction (IE) and later particularizes on OBIE, and the last subsection provides an overview about NLI.

2.2.1 Knowledge Representation

Research in KR includes studying how to reason accurately and effectively and how to use symbols to represent a set of facts within a knowledge domain. As defined by Sowa (2000), "knowledge representation is the application of logic and ontology to the task of constructing computable models for some domain". Davis et al. (1993) argue that this notion can best be understood in terms of five distinct roles that it plays: (1) a surrogate, (2) a set of ontological commitments, (3) a fragmentary theory of intelligent reasoning, (4) a medium for pragmatically efficient computation, (5) a medium of human expression. In general, KR implies creating surrogates that represent real world entities, and endow them with properties and interactions that represent real world properties and interactions.

Ontology is a central concept in KR. It is formally defined as an explicit specification of a shared conceptualization (Gruber, 1993). It describes a hierarchy of concepts related by subsumption relationships, and can include axioms to express other relationships between concepts and to constrain their intended interpretation. Guarino (1998) proposed a classification scheme for ontologies based on their level of generality: (1) top-level ontologies describe very general concepts like space, time, matter, object, event, action, etc., and are independent of a particular problem or domain; (2) domain ontologies / task ontologies describe, respectively, the vocabulary related to a generic domain like government, medicine, automobiles, or a generic task or activity like diagnosing or selling; (3) application ontologies describe concepts which often correspond to roles played by domain entities like temporary document or pending approval.

From the computer science point of view, the usage of an ontology to explicitly define the application domain brings large benefits regarding information accessibility, maintainability, and interoperability. The ontology formalizes and allows to make public the application's view of the world (Guarino, 1998). Ontologies allow specifying knowledge in machine processable formats. They are machine processable since they are specified using languages with well defined syntax such as Resource Description Framework Schema (RDFS) and Web Ontology

Language (OWL). Also, because ontology specification languages have well defined semantics, specifying knowledge using ontologies prevents the meaning of the knowledge to be open to subjective intuitions and different interpretations (Antoniou and van Harmelen, 2009).

Some research works produced mature, easy to use tools to create and maintain ontologies (Noy et al., 2000; Knublauch et al., 2004). Others focused in developing semantic reasoners. A semantic reasoner is a piece of software able to infer logical consequences from the set of asserted facts or axioms of the ontology. Many reasoners use first-order predicate logic to perform reasoning but there are also examples of probabilistic reasoners (Sirin and Parsia, 2004; Klinov, 2008).

2.2.2 Information Extraction

The general problem of Information Extraction (IE) involves the analysis of natural language texts to determine the semantic relations among the existing entities and the events they participate in: their relations. Informally, the goal is to detect elements such as "who" did "what" to "whom", "when" and "where" (Màrquez et al., 2008). Natural language texts can be unstructured, plain texts, and/or semi-structured machine-readable documents, with some kind of markup. The information to be retrieved are entities, classes of objects and events, and relationships between all them (Riloff, 1999). As Gaizauskas and Wilks (1998) observed, IE may be seen as populating structured information sources from unstructured, free text, information sources.

IE is different from information retrieval, which is the task usually performed by current search engines such as Google. IE aims to extract relevant information from documents, and information retrieval aims to retrieve relevant documents from collections (see Figure 2.2). In the case of the later, after querying search engines, users must read each document of the result set for knowing the facts reported. However, when the goal is to get a summary of facts reported in large amounts of documents, or having facts presented in tables, IE becomes a more relevant technology (McNaught and Black, 2006).

Two important challenges exist in IE. One derives from the variety of ways of expressing the same fact. As illustrated by McNaught and Black (2006), the next statements inform that a woman named Torretta is the new chair-person of a company named BNC Holdings:

- BNC Holdings Inc. named Ms. G. Torretta to succeed Mr. N. Andrews as its new chair-person.
- Nicholas Andrews was succeeded by Gina Torretta as chair-person of BNC Holdings Inc.
- Ms. Gina Torretta took the helm at BNC Holdings Inc. She succeeds Nick Andrews.

To extract the relevant information from each of these alternative formulations it is required linguistic analysis to cope with grammatical variation (active/passive), lexical vari-

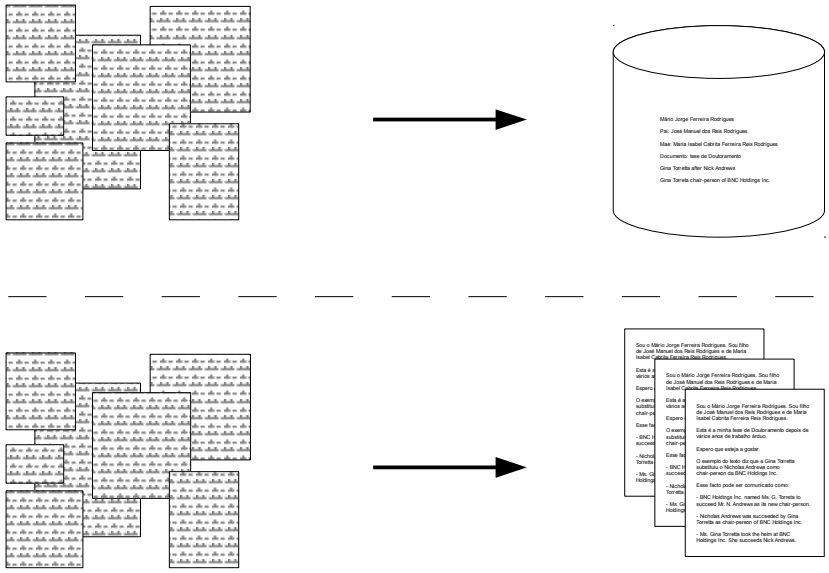


FIGURE 2.2: Difference between information extraction (above dashed line) and information retrieval (below dashed line). Image source: http://gate.ac.uk/ie/.

ation ("named to"/"took the helm"), and anaphora resolution for cross-sentence references ("Ms. Gina Torretta.... She...").

The other challenge, shared by almost all NLP tasks, derives from the high expressiveness of natural languages, which can have ambiguous structure and meaning. Lee (2004) exemplifies this phenomenon with a McDonnell-Douglas ad from 1985: "At last, a computer that understands you like your mother". This sentence can be interpreted in, at least, three different ways: (1) the computer understands you as well as your mother understands you; (2) the computer understands that you like your mother; (3) the computer understands you as well as it understands your mother.

Figure 2.3 illustrates part of the syntactic structure of the first two interpretations. It is clear that the structure of the sentence is directly related to its meaning. The challenge is to find which structure is the (most) correct one, knowing that the sentence context influences the meaning, and thus the most appropriate structure is not exclusively an intrinsic value of the sentence.

Over the years several different approaches have been proposed to solve the challenges of IE and of NLP in general. These approaches can be categorized as (1) rule based approaches which use hard coded rules encoding the solution for a given task. These systems are usually targeted for specific languages and domains, generally being ready to use and very accurate but difficult to port for other languages and/or domains; (2) machine learning approaches which use (annotated) corpora to train probabilistic models that learn how to perform a given task. These systems are usually adaptable to different information domains and/or languages and a major shortcoming is that, if not available, creating (annotated) corpus is a time consuming

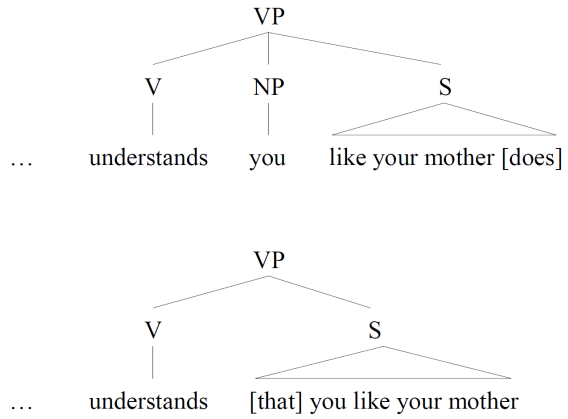


FIGURE 2.3: Two possible syntactic structures for the last part of the sentence "At last, a computer that understands you like your mother".

task that should be performed by specialists in the area; (3) hybrid approaches which use a mix the previous two. The aim is combine the best features of each kind of approach: the accuracy of rule based approaches with the coverage and adaptability of machine learning approaches.

Some IE approaches use ontology to store and guide the IE process. The success of these approaches motivated the creation of the term Ontology-Based Information Extraction (OBIE). The next subsection will elaborate on OBIE as it is a relevant concept for this work.

Ontology-Based Information Extraction

Wimalasuriya and Dou (2010) define OBIE as a "system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies." Different approaches to IE and, in particular, to OBIE have being proposed and developed over the years. The approaches can differ in some dimensions including (the first two dimensions are relative to IE in general):

- The identification and extraction of information can be performed using probabilistic methods or explicit defined sets of rules.
- The types of document from which information is extracted can be unstructured, plain text, or semi-structured and structured sources.
- The ontology can be constructed from the documents content or exist before the beginning of the extraction process. In both cases the ontology structure can be updated automatically while processing documents or not.

- The kind of information extracted varies from extracting only ontological instances to extracting ontological classes and properties. These different approaches are correlated with the previous point.

A frequent approach to extract information about generic domains is to use Wikipedia to build their knowledge base (Bizer et al., 2009; Suchanek et al., 2007; Wu et al., 2008). Wikipedia structure is used to infer the semantics, and the knowledge base is populated by extracting information from pages texts and infoboxes. The approaches that do not take advantage of Wikipedia structure acquire information from generic web pages. The knowledge base structure is often inferred from pages content and the knowledge base is populated using the same sources (Etzioni et al., 2004; Yates et al., 2007).

Most approaches, whether using Wikipedia or not, use shallow linguistic analysis to detect the information to extract. Shallow analysis involve detecting text patterns and, at most, using part-of-speech information: which words are nouns, verbs, adjectives, etc. The use of shallow linguistic information makes difficult the acquisition of information from complex sentences. The most relevant systems and respective approach is summarized in Table 2.3.

2.2.3 Natural Language Interfaces

Natural language interfaces aim to allow people to interact with machines using a human language, such as English or Portuguese, as opposed to a computer language, command line interface, or graphical user interface. The interaction can be either written and/or spoken and, as human interaction is based on conversation between persons, natural language interfaces try to emulate the conversational aspects of human interactions such as turn-taking and speech acts. Advances in automatic speech recognition, language understanding, language generation, and speech synthesis enabled the emergence of complex conversational language interfaces (Jurafsky and Martin, 2008; Bohus and Rudnicky, 2009; Bohus, 2013).

Today NLI are not able to process full natural languages. NLI mainly use what is called Controlled Natural Language (CNL). As defined by Kuhn (2013), a "controlled natural language is a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax and/or semantics while preserving most of its natural properties". Schwitter (2002) differentiates CNL according to the problem they are supposed to solve: (1) improve communication among humans, especially speakers with different native languages; (2) improve manual, computer-aided, semi-automatic, or automatic translation; (3) provide a natural and intuitive representation for formal notations. Here the focus will be on the third type of language as it is the most relevant for this work.

Kuhn (2013) reports some studies testing the usability of CNL tools and others evaluating the comprehensibility of the actual languages. Regarding the usability studies, the general conclusion was that usability is improved by using CNL. Regarding the comprehensibility studies, if common users are able to correctly interpret given statements, CNL proved easier

Table 2.3: Comparison of ontology based information extraction systems. Source (Wimalasuriya and Dou, 2010).

System	Information Extraction Method	Ontology Construction and Update	Elements Extracted from Texts	Sources
Embley (Embley, 2004)	Linguistic rules	Pre-built Not updated	Instances, property values	Documents of a domain
iDocuments (Adrian et al., 2009)	Linguistic rules, gazetteer lists	Pre-built Not updated	Instances, property values	Documents of a domain
KIM (Popov et al., 2004)	Linguistic rules, gazetteer lists	Pre-built Not updated	Instances, property values	Documents of a domain
Kylin (Wu et al., 2008)	Classification, Web search	Runtime Not updated	Classes, taxonomy, properties, instances	Wikipedia
OntoSyphon (McDowell and Cafarella, 2006)	Web search	Pre-built Not updated	Instances	Any
OntoX (Yildiz and Miksch, 2007)	Linguistic rules	Pre-built Not updated	Instances, property values	Documents of a domain
PANKOW (Cimiano et al., 2004)	Web search	Pre-built Not updated	Instances	Any
SOBA (Buitelaar et al., 2006)	Linguistic rules, gazetteer lists	Pre-built Not updated	Instances, property values	html files of a domain
Vulcain (Todirascu et al., 2002)	Partial parse trees	Pre-built Not updated	Instances, property values	emails of a domain
Maedche et al. (2003)	Partial parse trees	Runtime Updated	Classes, taxonomy, properties, instances	Documents of a domain
Saggion et al. (2007)	Linguistic rules, gazetteer lists	Pre-built Not updated	Instances, property values	Documents of a domain

and faster to understand than a common ontology notation. Experiments that gave mixed results are also reported.

Damljanovic and Bontcheva (2009) present an overview of NLI to knowledge bases. They found that users behave differently when communicating with computers than with humans. In the latter case, their conversation relies heavily on context. In the former case, users tend to restrict the vocabulary as they make assumptions about what computers can and cannot understand. This means that users automatically try to adapt to the CNL of the system. However this brings the problem of habitability. Habitability refers to how easily, naturally and effectively users can use language to express themselves within the constraints imposed by

the system. If users can express everything they need for their tasks the system is considered habitable (Ogden and Bernick, 1997). Another way of viewing habitability is as the matching between users' expectations and the capabilities of NLI systems (Bernstein and Kaufmann, 2006).

2.3 Semantics and HLT in e-Government

The usage of semantic and HLT in e-government can bring several important benefits (Rodrigues et al., 2010b). It allows access to unstructured information that exists in natural language documents. Relevant sources of information for e-government are originally created in natural language documents, such as forms, laws, regulations, etc. Those documents are usually available to the public in Portable Document Format (PDF) and HyperText Markup Language (HTML) formats. A computer system that does not feature NLP capacities cannot derive meaningful information from it, being limited to store and display. Systems able to understand the information of this type of documents could assist users more efficiently by selecting the information that is relevant in the usage context. It also makes possible developing NLI which allows, in the e-government context:

- More intuitive user interfaces - natural language interfaces let people express themselves in their own language, not forcing individuals to relearn how to communicate. Profiting from the recent advances of natural language interfaces in other areas, it is possible to develop e-government support systems that are able to solve user queries formulated in natural language, instead of forcing users to look for instructions or to find information in the set of rules that regulate the service. A frequent complaint against natural language interfaces is that the linguistic coverage is not obvious, in other words users do not know what to say. This problem has been addressed and there are strategies that allow dialog systems to intuitively drive the user to use the right vocabulary set (Gorin et al., 1997).

- Dialog becomes an option - the use of speech and natural language allows a solution to be achieved by iteratively specifying a problem using dialog. Dialog processing is, by its very nature, incremental. No dialog agent (artificial or natural) processes whole dialogs, if only for the simple reason that dialogs are created incrementally, by participants taking turns. An incremental system can work with units smaller than utterances, allowing the creation of a more reactive system capable of taking the initiative in the conversation to clarify, ask for missing information, or suggest. By having a conversation, individuals that do not know which specific service solves their problem would be able to explain it to discover the solution, instead of having to make a thorough search to find how they can be served.

- Take advantage of voice channels - people are allowed to use their telephone and have

the same level of service they would have using the web. Speech is the most natural and easy existent interface, not only for people with special needs, but for people in general, as Nass and Brave (2005) state: "Ubiquitous computing - access to all information for anyone, anywhere, at any time - relies on speech for those whose eyes or hands are directed to other tasks (such as driving ...) or for those who cannot read or type (such as children, the blind, or the disabled)". Another advantage is that telephones are more popular than computers and more people feel comfortable with them. Automated services provided by voice are already available to the general public (e.g. Google Voice).

- Access for all - the combination of the above advantages contributes to reduce the digital divide. The usage of written natural language and speech interfaces facilitates the access of minorities as the visual impaired, people with severe speech disabilities, and people not familiar with ICT and/or with low literacy levels. The usage of ubiquitous telephone or smartphone allows services to become virtually accessible from anywhere and at any time, increasing the potential to reach more users, including those who live in remote areas, are homebound, etc. Written interfaces are usually considered robust to be used by the general public, while speech technology is often considered not ready for massive use. However, today speech technology is already in a quite advanced and mature stage of development, which can be demonstrated by the existence of several commercial products in different languages.

Conventional methods to catalog and find information (e.g. search engines) have two shortcomings when applied to e-government: first, they often produce a large number of results to a query, relying on the client's ability to select the appropriate ones; second, they do not integrate related information scattered across documents. Government integrated workflows frequently depend upon multiple services contributing for a single outcome (e.g. building a house may depend on services provided by several government agencies at local, regional or national level). The lack of information integration causes results to be a collection of fragments to be compounded in a meaningful way by clients. This composition should be performed by an entity familiar with the subject because clients shouldn't have to spend time learning how services work to understand the outcome (Paiva Dias and Rafael, 2007; Rodrigues et al., 2010b).

Some works were dedicated in studying how semantic and HLT could improve e-government information provision. An ontology based information system named EgoIR used a specific ontology to associate concept instances with documents (Ortiz-Rodríguez et al., 2007). In this work, associations are done manually and the authors claim that the queries returned documents more accurately. No usage examples were demonstrated and the system did not processed or extracted information inside the documents.

Iriberri and Leroy (2007) developed a crime reporting system featuring NLP technologies. NLP was based on POS tags, gazetteers, and semantic tagging with handcrafted rules ana-

lyzing witnesses' narratives to extract the information that was required in standard police reports. The system was developed for police reports domain only.

In project HOPS (Gatius et al., 2006) the approach was to develop a multilingual dialogue system to guide users when accessing local administration public web content. It blended voice and text channels and the problem of inserting data into the system was not specifically addressed.

Another work developed an experimental website, `Commentonthis.com`, designed to enable having detailed discussions around the contents of major public documents. The website allows citizens to share their views on the details of some government documents, which have been split into paragraphs in order to make them "commentable". Other examples of works aimed at covering the semantic e-government domain are: the DIP project[1], the Reimdoc project[2], The IFIP Working Group 8.5[3], the Ontogov project[4], the Egov project[5], and the WEBOCRAT project[6] (Gómez-Pérez et al., 2005).

Other e-government projects used semantic technologies to specify, develop, and deploy services, and were rather centered in solving problems as interoperability and service integration. Examples of such projects are OneStopGov (Chatzidimitriou and Koumpis, 2008) and Access-eGov (Sroga, 2008).

Some governments already defined standards respecting semantic interoperability. The goal of these initiatives is allowing the community to develop systems able to communicate with other using the defined standards. Table 2.4 presents some of the most mature works.

2.4 Summary

This chapter reviewed the current challenges and recent relevant works in some subareas of e-government, semantics and HLT, and on the crossing of these broad areas.

One prominent trend in e-Government is studying how to take advantage of Web 2.0 technologies to improve communication between public officials and society: citizens, enterprises, and other branches of government. The main idea is that government should be a platform of data provided in formats suitable to be automatically processed by third party systems as well as consumed by people. This trend is called e-government 2.0.

Relative to semantic and HLT research, one challenge being tackled recently is how to perform IE at the web scale. This brings challenges in how to interpret a wide variety of documents and how to build knowledge representations for the multitude of concepts found in those documents. Another topic reviewed in this chapter is NLI, particularly using CNL.

[1] http://dip.semanticweb.org
[2] http://reimdoc.atosorigin.es
[3] http://falcon.ifs.uni-linz.ac.at/research/ifip85.html#aim
[4] http://www.ontogov.com/
[5] http://www.egov-project.org
[6] http://www.webocrat.org/

Table 2.4: Semantic interoperability standard definition efforts lead by governments.

Name	Technology and purpose
oeGov, US	OWL ontologies to enable distributed creation and maintenance of metadata about government data.
govtalk, UK	Definition of the elements, refinements and encoding schemes to be used by government.
ADAE, France	Reference schema, core component, category, or semantic asset.
InfostructureBase, Denmark	Repository of XML schemas in the Danish Public Sector about business process descriptions, datamodel descriptions, interface descriptions, and XML schemas from public and private organizations.
SAGA, Germany	Repository of XML Schemas for documents and a set of codelists.
HKSARG, Hong Kong	XML Schema Design and Management Guide.
Joinup, EU	XML and RDF to find, choose, re-use, develop, and implement open source software and semantic interoperability assets.
NIEM, US	Partnership between the U.S. Department of Justice and the Department of Homeland Security to develop, disseminate and support enterprise-wide information exchange standards and processes.

These interfaces help to improve people access to those big sources of information by allowing a higher level of expressiveness than keyword based search.

On the crossing of these areas, e-government and semantic and HLT, is still necessary more research and development. Several important initiatives are already deployed to society. As these initiatives are usually held by central governments it is necessary to provide tools to support them and to allow other tiers of government (regional, local) to also deploy such initiatives.

3

Conceptual Model

The two previous chapters described relevant e-government opportunities and challenges; discussed current approaches and research projects in e-government; and presented technologies that have the potential to improve e-government information delivery.

Based on these previous considerations, it is now presented a conceptual model for government information provision. The model aims to make effectively available relevant information of government documents written in natural language. This implies the ability to: interpret natural language texts; discriminate which information is relevant; be accessible anytime and anywhere; and provide a suitable access to all citizens including those with low ICT skills. In order to have a tool that can be realistically adopted by government services, it is also necessary to support an easy and seamless adaptation to the information domain. Alongside these features, a main concern was to keep the model simple and clear.

This chapter is organized in five sections. The first section introduces the view of the future used as guideline in the model design. The second section discusses the model requirements based on the future view and other concerns. The third section presents the model and explains how it satisfies all requirements enumerated. The fourth section outlines three usage scenarios that show different contexts were the model can be used. The fifth section closes the chapter with a summary.

3.1 A View of the Future of e-Government

E-Government evolution over the coming years will be influenced by technological developments as well as by contextual factors, such as social and economic trends. A study requested by the Institute for Prospective Technological Studies, a scientific institute of the European Commission's Joint Research Centre, took the long view to provide input for e-government longer-term strategic planning in the European Union (Frissen et al., 2007). It did so by identifying emerging trends and opportunities for enhancing governments and governance in 2020.

The study took into account a wide array of technological, social, economic and political factors that can condition government and governance challenges in the future. These different factors were organized in four political scenarios ranging from high citizen engagement to low citizen engagement, and from heterogeneous European culture to homogeneous European culture. Each scenario described a coherent and consistent picture of the future based on possible developments in society, politics, the institutional sphere, the economy, and in the technological domain. The scenarios are briefly explained in Table 3.1, and Figure 3.1 shows how the scenarios are positioned in terms of citizen engagement and type of culture.

Table 3.1: Characterization of the political scenarios outlined by Frissen et al. (2007). Source (Frissen et al., 2007, p. 57).

Scenario	Characterization	Explanation
Our Europe	Utopia	Governments exert control, citizens have a balanced view on privacy; politics may seem in control but ample means of engagement for citizens and business ensure a healthy interaction on the key themes in society.
We, The Market	Background Government	Market parties are in the lead; government is facilitator and needs to prove its legitimacy. The companies self-proclaimed stewardship of public data demonstrates the take-over by private parties of public concerns. In the name of economic growth and job security citizens endure interference in private matters and exclusion.
My Community	Fluid Government	Everything is in flux. Government co-operates with market parties, but each activity needs to be negotiated with changing stakeholder groups. There is a continuous threat that events determine the political agenda and block efficient governance.
Me, Myself and I	Government's Sweat Shop	Even more than in the previous scenario, government has to express its legitimacy and has to counter the well-determined self-interest of the population. Politics is ending up in a one-to-one power game between public authorities and individual citizens or non-governmental organizations.

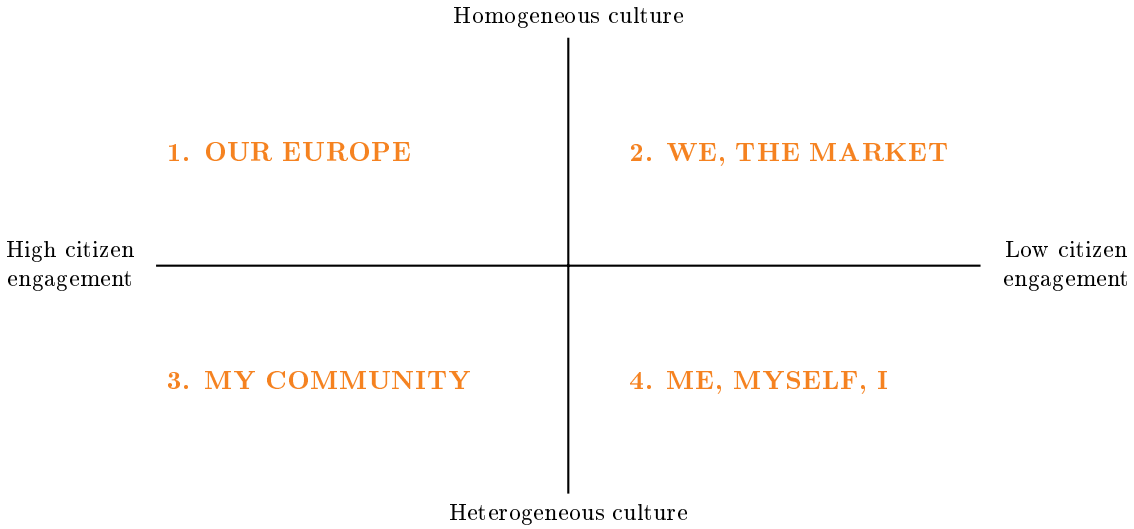


FIGURE 3.1: Political scenarios outlined by Frissen et al. (2007): citizen engagement - extent to which citizens and companies are involved in political and societal processes; culture type - extent to which citizens of Europe agree on the role and future of Europe and the extent to which they share norms and values. Source (Frissen et al., 2007, p. 52).

Considering all scenarios, the authors of the study foresee empowerment of citizen and enterprises as the trend to have the most significant impact in the coming decades. As a consequence, "Governments will have to operate in more open and networked constellations with other stakeholders. They will need to find the balance between being extremely transparent and accountable, on the one hand, and operate in a flexible, not overtly bureaucratic way on the other hand. They will need to act as intelligent, all-knowing government and deliver services that are highly sophisticated, personalized and pro-active..." (Frissen et al., 2007, p. 109). The report pinpoints technological challenges in achieving this goal:

- Ensure technological interoperability and standardization.
- A stronger investment in technologies that enable smart ways of cooperating and sharing or producing knowledge ("collective intelligence", open source and open content, collaborative computing tools, etc).
- Stimulate the use of technologies which are designed to cope with potential information overload.
- Ensure that networks and services are accessible to all, both on the level of infrastructures, as on the level of services and the necessary interfaces.
- Reduce the dependency on ICT infrastructures and related services or build the necessary safeguards.

- Governmental transformation requires back office reorganization and one-stop shop approaches, which, in turn, require substantial process and workflow redesign that needs to be translated into new information architectures. An additional challenge is that these new architectures need to be flexible and open in order to be sufficiently user-centered and dynamic.

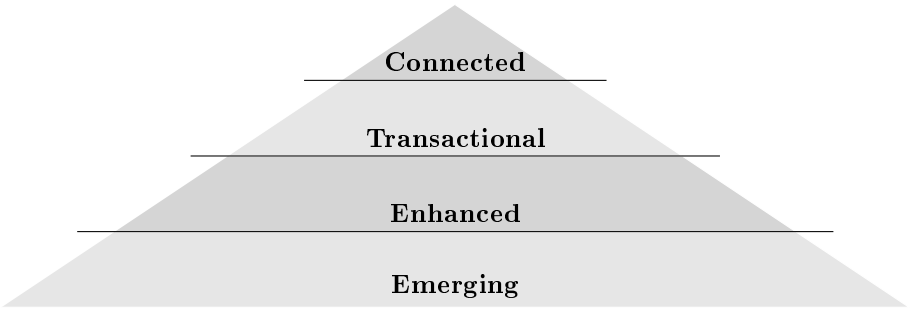
In the present day, these technological challenges are reflected in the European eGovernment Action Plan 2011-2015 (European Commission, 2010). The Action Plan is structured along the four main priority areas of the Malmö Declaration (Malmö Declaration, 2009) which called for actions contributing to a new generation of open, flexible and personalized e-government services of administrations at local, regional, national, and European level. The first enunciated priority area is User Empowerment and the milestones to focus on are: better access to public sector information; transparency of governments and public administrations; and effective means enabling the active involvement of citizens and businesses in the policy-making process based on newly available technologies. The other three priority areas of the Malmö Declaration are: Internal Markets; Efficiency and Effectiveness of Governments and Administrations; Preconditions for eGovernment.

This perspective of the near future - where should be addressed concerns as transparency, information overload, effective access to information, etc. - is not exclusive to the European Union. The 2010 edition of the United Nations E-Government Survey (United Nations, 2010) shares a future view that, although not assumed as the United Nations official view, is the only one presented: the idea of "government 2.0" associated with the use of social networks by the public sector should assume a broader definition - government should be viewed as a platform. From this perspective, government should become a provider of data and services that others can also exploit as they see fit, allowing third parties to innovate by building upon government data and applications.

However, in the present day, most governments are not yet sharing information using open data. Open data is about making information freely available to everyone, without restrictions from copyright or patents and in standard machine-readable formats that can be exploited without the use of any given piece of software. In (United Nations, 2010, p. 16) it is stated that "open data enhances public sector efficiency by transferring some of the analytical demands of government to third parties such as non-governmental organizations, research institutes and the media, which have been found to combine data from various sources in original and inventive ways". The belief is that if governments provide data in a non-proprietary and predictable format, third parties are likely to maximize the value of the information, hence contributing to provide services that better respond to users' expectations and needs. Thus, third parties can play an important role in the co-provision of services of high public value by deploying technologies in a manner that is creative and innovative. This can be stimulated using open data principles since they allow reducing the entry barriers for non-governmental

parties.

The United Nations Public Administration Programme created a web measure index to assess the maturity level of e-government initiatives around the world. The index is based on a four stage model that defines the stages of e-government development according to scale of progressively sophisticated citizen services (see Figure 3.2). As countries progress, they are ranked higher in the model according to a numerical classification corresponding to the four stages.



Stage 1 **Emerging information services** Government websites provide information on public policy, governance, laws, regulations, relevant documentation and types of government services provided. They have links to ministries, departments and other branches of government. Citizens are easily able to obtain information on what is new in the national government and ministries and can follow links to archived information.
Stage 2 **Enhanced information services** Government websites deliver enhanced one-way or simple two-way e-communication between government and citizen, such as downloadable forms for government services and applications. The sites have audio and video capabilities and are multi-lingual. Some limited e-services enable citizens to submit requests for non-electronic forms or personal information, which will be mailed to their house.
Stage 3 **Transactional services** Government websites engage in two-way communication with their citizens, including requesting and receiving inputs on government policies, programmes, regulations, etc. Some form of electronic authentication of the citizen's identity is required to successfully complete the exchange. Government websites process non-financial transactions, e.g. e-voting, downloading and uploading forms, filing taxes online or applying for certificates, licenses and permits. They also handle financial transactions, i.e. where money is transferred on a secure network to government.
Stage 4 **Connected services** Government websites have changed the way governments communicate with their citizens. They are proactive in requesting information and opinions from the citizens using Web 2.0 and other interactive tools. E-services and e-solutions cut across the departments and ministries in a seamless manner. Information, data and knowledge is transferred from government agencies through integrated applications. Governments have moved from a government-centric to a citizen-centric approach, where e-services are targeted to citizens through life cycle events and segmented groups to provide tailor-made services. Governments create an environment that empowers citizens to be more involved with government activities to have a voice in decision-making.

Figure 3.2: United Nations Public Administration Programme four stage model to measure e-government evolution. Source United Nations (2010, p. 95).

From the model it is observable that e-government's first concern (stage 1) should be

information provision: websites that provide information on public policy, laws, regulations, etc. Stage 2 further elaborates on these concerns and, in the most sophisticated stage (stage 4) one goal is to transfer information, data and knowledge from government agencies using integrated applications. This goal implies the existence of open data.

In this dissertation, the view to the future adopted is the one shared by both European Union and United Nations. This view is obviously shared by the majority of, if not all, countries that participate on those institutions. Resuming, an important part of the future of e-government is about:

- Making government information available. This is important today and should be important in the future even considering a wide range of political scenarios.
- Avoiding information overload. Government produces large amounts of documentation and it is important to provide useful information and not flood citizens and enterprises with random data.
- Providing data in machine-readable formats using open standards. This is necessary to involve third parties in co-provisioning government services and information. Besides the potential to help government achieving its mission, involving third parties increases data and infrastructure redundancy, thus adding safeguards and reducing the dependency on government ICT infrastructure.
- Making information accessible to all citizens and enterprises. This is very important to stimulate e-government appropriation by the community, and implies among other things the usage of friendly and intuitive user interfaces.

3.2 Requirements

This section enumerates the general requirements for the proposed conceptual model. The requirements were identified based on the view of the future presented in the previous section and on generic concerns regarding information provision. Each of the following five subsections elaborates on a requirement: information acquisition from natural language documents, data storage using a knowledge base, relevant information identification, interoperability, and user interface.

3.2.1 Information Acquisition from Natural Language Documents

Important sources of government information are originally created in natural language plain text documents such as laws, regulations, etc (Rodrigues et al., 2010b). Thus a requirement for the model is to acquire information from natural language texts. Acquiring information from such documents improves efficiency because it avoids having government services

replicating content in various formats. However, this requirement brings some technological challenges. Conventional methods to catalog and find textual information (e.g. search engines) have two shortcomings when applied to e-government: (1) they often produce a large number of results to a query; (2) they do not integrate related information scattered across documents. Government integrated workflows frequently depend upon multiple services contributing for a single outcome.

Both shortcomings can be mitigated if a system possesses the ability to analyze and understand texts. This task is usually called Information Extraction (IE) (see Section 2.2.2 for details) and it implies having a conceptual model that supports NLP algorithms. It is also necessary to integrate and store extracted information is a suitable format. This will be addressed in the next requirement.

As an addition to this requirement, the model should also be able to obtain information from external structured sources. This would allow complementing the government's information with external data. An example is a municipal report about streets under repair. Alternative to showing a list of street names, this information can be displayed in a map if the coordinates are obtained from a location server as OpenStreetMap[1].

3.2.2 Data Storage using a Knowledge Base

According to Akerkar and Sajja (2010), a Knowledge Base (KB) is an information repository that includes utilities to facilitate knowledge retrieval, learning and justification, and transfer expertise from one domain of knowledge to another. It is a structured framework that has the ability to organize information by its meaning. They support information integration and are becoming an increasingly important part in web and enterprise search (Bizer et al., 2009).

As referred in the previous requirement, information about a topic can be scattered across documents since government workflows frequently depend upon multiple services. In these cases, the overall picture is just obtained after compounding the information fragments in a meaningful way. This composition should be done before providing the information to clients. Clients should not have to spend time learning how services work to understand the outcome.

As KBs usually conform to ontologies, the requirement for the model is to feature a KB that conforms to domain ontology, in this case e-government domain (see Section 2.2.1 for details about KBs, ontologies and respective classification). An important feature of KBs is the possibility of performing logic inference. Based on ontology, context-aware computing can exploit various existing logic reasoning mechanisms to infer new facts from already known ones. It is also possible to check and solve inconsistent knowledge that can occur due to imperfect knowledge acquisition (Wang et al., 2004). Other benefits will be elaborated in the next requirements such as improvement of information accessibility, maintainability, and

[1] http://www.openstreetmap.org

interoperability by having a formalized and public application's view of the world (Guarino, 1998).

3.2.3 Relevant Information Identification

Government services produce large amounts of documentation. Disclosing such large volumes can create a noise effect and becomes a way to hide information. Moreover, if not properly addressed, information disclosure can require a lot of time and effort, and can be hard to keep up with all the demands (Florini, 1999). Surveys across Europe (Colclough and Tinholt, 2009), United States (Reddick, 2005) and other countries (Andersen et al., 2010; Dias, 2011) show that it's still necessary to improve the capacity of providing government's relevant information to clients and not flood them with random data. This implies the ability to locate and discriminate the relevant parts of the existing information.

A difficulty of this requirement is that information relevancy is not an absolute value. It depends on several factors as: the purpose of the document (law, an informal comment on some topic, etc), the core business of the publisher (for which information the publisher is an authority and which information serves as complement), and the expertise of the target audience on a given subject (what is the appropriate level of detail). Also, due to changes in laws, regulations and society needs, information relevant today can be irrelevant tomorrow and vice-versa. These factors make inappropriate to define, *a priori* and for all government services, the set of information that is (more) relevant in a given context, the same is to say the information domain.

A solution involves having the ability to adapt the information domain as needed. Adapting the information domain means to accept changes in an easily and timely manner, in order to reduce costs and speed up processes. It also means to be able to locate and acquire relevant information for the new domain, doing so without significant system reconfiguration or setup. It is important to take this feature into account from the beginning because it is usually cost intensive to keep knowledge bases up-to-date as the domain changes (Bizer et al., 2009).

3.2.4 Interoperability

The Institute of Electrical and Electronics Engineers (IEEE) defines interoperability as the "ability of a system or a product to work with other systems or products without special effort on the part of the customer. Interoperability is made possible by the implementation of standards." (IEEE, 2010). Interoperability is based on agreements between requesters and providers on message passing protocols, procedure names, error codes, etc. Semantic interoperability ensures that these exchanges make sense: that the requester and the provider have a common understanding of the meanings of the requested services and data (Heiler, 1995). Current and future views of e-government urge its information systems to work in

networked environments making semantic interoperability an important feature to take into account in the proposed conceptual model.

The proposed model needs to provide data using open standards complying this way with the interoperability requirement. As ontologies formalize the application's view of the world, semantic interoperability can be achieved by defining and making public the ontology using an open standard (see Section 2.2.1 for details about ontology standards). This is a contribution for the Semantic Web. Semantic Web is about web applications that facilitate information sharing, interoperability, and collaboration on the World Wide Web (WWW) (Berners-Lee et al., 2001).

Ontologies facilitate knowledge sharing because they define the semantic context. The use of context enables computational entities to have a common set of concepts about context while interacting with each other. They also facilitate knowledge reusing. By reusing well defined web ontologies of different domains like temporal and spatial ontology, it is possible to compose large-scale context ontology without starting from scratch (Wang et al., 2004).

3.2.5 User Interface

Government must serve all its clients regardless of who they are, where they are located, or of their specific peculiarities (Rodrigues et al., 2010b). Digital divide is an important concern in any context and in e-government in particular. If not properly addressed, online service provision can contribute to increase this problem (see Section 2.1.2 for details). A central problem faced by would be users is that of formulating queries in terms that are communicable to the system.

Database query languages can be intimidating to the non-expert. Supporting arbitrary natural language queries is regarded by many as the ultimate goal for a query interface (Li et al., 2005). Natural language interfaces let people express themselves in their own language, not forcing individuals to relearn how to communicate. The interaction can be spoken or written. Written interfaces are usually considered robust to be used by the general public, while speech technology is often considered not ready for massive use. So current applications should provide, at least, a written interface but can provide both. In fact, supporting both written and spoken interfaces brings several advantages. The usage of written and speech interfaces combined facilitates the access of minorities as the visual impaired, people with severe speech disabilities, and people not familiar with ICT and/or with low literacy levels (see Section 2.2.3 for details). The usage of ubiquitous telephone or smartphone allows services to become virtually accessible from anywhere at any time, increasing the potential to reach more users, including those who live in remote areas, are homebound, etc (Rodrigues et al., 2010b).

Accepting queries formulated in natural language (written and/or spoken) is the goal for this requirement. Regarding information output, it should be concise and have a reference

to the document(s) where the data was found. It is important to trace all data to the legal documents and, this way, clients are not overwhelmed with lots of data, having access to more information if desired. Alternative modes of presentation should be allowed when appropriate (tables, maps, etc).

3.3 Model Overview

In this section is presented a model that complies with the requirements enumerated in the previous section. Its main advantage is to provide a simple and clear outline of the overall system operation (see Section 2.2.2 for other proposals). It is an original proposal that clearly separates the realms of (natural) language and (semantic) knowledge.

The model is organized in three main components. The first component accomplishes with the `Information Acquisition from Natural Language Documents` requirement and is named Natural Language Processing (NLP) by being based on technologies from the NLP area. Its goal is to obtain structured data from natural language unstructured texts. For this, it includes tools to obtain documents from the WWW and/or from local file systems and extract their content. Then, it takes these texts as input and produce fixed-format, unambiguous data, as output. The output can be more or less sophisticated but it needs to have enough detail to discriminate sentences with different meanings. This depends on the type of texts to be processed and on application purposes. Taking into account the state of the art in NLP, it is expectable to have an output that identifies named entities and have some kind of syntactic structure associated. This should be the only component of the model to be reprogrammed when changing or adding another natural language.

The Knowledge Representation (KR) component is the second one, and contributes to the `Relevant Information Identification` requirement. Its purpose is to bridge the gap between the realms of natural language and knowledge representation. In order to do so, it contains tools to define data semantics and associate it with output samples of the natural language processing component. The data semantics depends on the application purposes and, considering the state of the art in KR and the enumerated requirements, it should be defined via a domain ontology, in this case for government. The ontology must also support references to information sources to ensure traceability and a concise output due to the `User Interface` requirement. It is advisable to encode references to the original documents using DCMI Terms[2] created and maintained by the Dublin Core Metadata Initiative (DCMI) because it is a *de facto* standard. Alongside the ontology definition, it is necessary to associate ontological classes and relations with output samples of the NLP component. These associations will later be used as learning examples by the SEI component. The outputs of this component are the ontology and the associations of ontology classes and relations to the NLP component output.

[2] http://dublincore.org/

The third component is named Semantic Extraction and Integration (SEI) and contributes to the requirements `Relevant Information Identification` and `Data Storage using a Knowledge Base`. Its task is to extract, organize, and store the output of the NLP component according to the semantic defined by the ontology. It needs to feature machine learning algorithms in order to support domain adaptation: the idea that a change in the information domain cannot cause a software reprogramming/re-engineering implies that the software adapts to, learns, the domain specification. These algorithms learn the associations provided by KR component and produce semantic extraction models. These models are used in runtime to associate semantic information from all data processed by the NLP component. Missing information - according to the ontology - can be searched in external structured sources. This search is conducted via structured sources connectors. The existence of structure in the external data makes possible to develop connectors that directly assign the appropriate semantics to the data. All information is stored in a knowledge base that conforms to the defined ontology.

The model also features interfaces for the semantic web and for human users. The semantic web interface addresses the `Interoperability` requirement. Data input is performed using structured sourced connectors just like the SEI component connector. Data output should be made providing data and its semantic. This implies to make the ontology public to let third party systems to know the system semantics. Data can be provided via web services, via query endpoints (SPARQL, Semantic Web Rule Language (SWRL), etc.), or via other standard mechanisms. The human interface addresses the `User Interface` requirement. It should support natural language input in written, spoken or both forms. The interface needs to dynamically adapt to the information domain due to the requirement of not having significant system reconfiguration when the information domain changes. This implies using interfaces that are aware of the system ontology and the knowledge base content. Figure 3.3 depicts the proposal.

3.4 Usage Scenarios

This section outlines usage scenarios for the proposed conceptual model. The scenarios were created to illustrate how government services, citizens, and enterprises benefit from information systems that comply with the proposed conceptual model.

Although this model is applicable in e-government at the central, regional and local levels, the discussion will focus at the local level. There are three reasons for this: (1) by narrowing possible applications, it clarifies the discussion by focusing in the information system itself and not so much in the deployment environment; (2) there is no loss of generality because it will not be addressed challenges exclusive to local government or avoided relevant problems of the other government levels; (3) if it was necessary to choose one government level in detriment of the others, the choice would fall on local e-government. The reason for this hypothetical choice

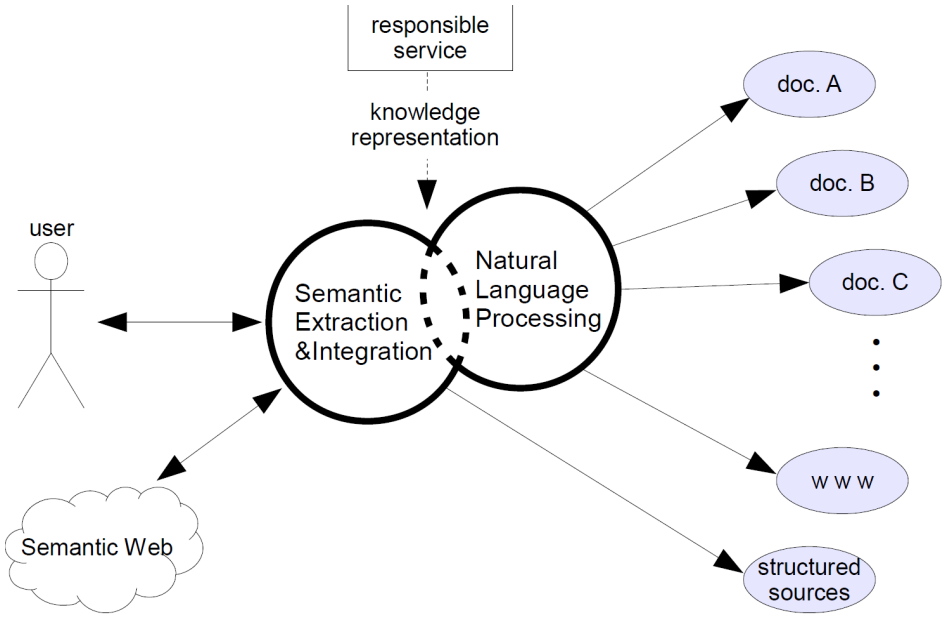


FIGURE 3.3: Proposed conceptual model: (natural) language and (semantic) knowledge are coupled according to the knowledge representation defined by the service responsible for the system. Single sided arrows represent data acquisition from natural language documents and from structured sources. Double sided arrows represent interfaces for the semantic web and for humans.

is that municipalities are often the closest point of service for citizens and enterprises (Rodrigues et al., 2010b). Also, evidence gathered in the 2010 European e-government benchmark survey show that local and regional e-government services are lagging behind when compared to central government (Lörincz et al., 2010). Figure 3.4 shows e-government maturity differences of four services across the tiers of government. Corroborating these evidences, Dias (2011) conducted a study for municipalities of Portugal, a country that has good classifications in international e-government benchmarks at national level services. The study found that despite the investment made in the last decades in e-government, Portuguese municipalities still exhibit medium level development in what relates information dissemination through the web.

Distinct types of entities can manage this model, that is to say essentially to be the responsible service that defines the data semantics and provide examples of data conforming that semantic. The usage scenarios are organized in three categories: managed by the information producer; managed by the information consumer; and managed by a third party. These scenarios are not mutually exclusive since the real setup can be a combination of them. Each of the next three subsections elaborates on one scenario.

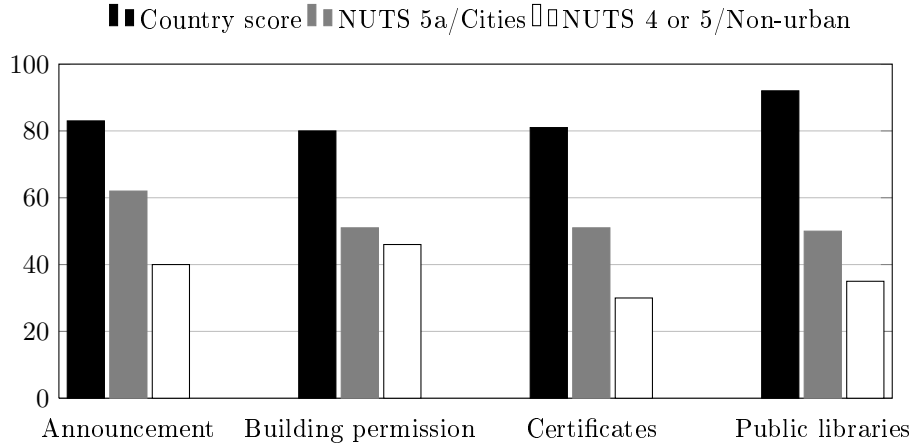


FIGURE 3.4: Four examples of service provision sophistication at three government levels: country level (country score); city level (NUTS 5a/Cities); and regional level (NUTS 4 or 5/Non-urban). Source Lörincz et al. (2010, p. 11). Nomenclature of Territorial Units for Statistics (NUTS) 4 are wide (regional) areas and was defined for most, but not all European Union (EU) Member States. NUTS 5 are municipalities or equivalent units in the 27 EU Member States.

3.4.1 Managed by an Information Producer

In this scenario, government services use the system to identify and extract semantic information from their documents. The ability of automatically capture and associate pieces of related data makes the provided information more complete, becoming easier to understand what government is doing and why. Also, it is possible to target information for specific audiences by defining which information is relevant for a given public. This customization helps to prevent information overload by allowing changing information priority according to the audience type.

This scenario will be now discussed for municipalities with an example that illustrates how citizens, businesses, and the municipality itself can benefit from this scenario. Consider a municipality that wants to keep the community informed about actions relative to city's infrastructure construction and maintenance. The benefit of automatically capture and associate related data is obtained by defining, in the ontology, that infrastructure construction and maintenance has the following parts: urban plan, architectural plan, land expropriation, construction budget, open call for constructors, etc. This way, knowledge base inference mechanisms will associate all information acquired about one of those topics with infrastructure construction and maintenance given that the same identifier is provided. Also, as different services contribute to urban plan, construction budget, etc, the system would learn, from given examples, how each involved service outputs its data thus preventing services to change methodologies when contributing for this common goal.

Regarding information customization, it is possible to setup profiles for citizens and busi-

nesses, and also to use a generic query entry point for more experienced users, whether using natural language or some query language. All profiles would use the generic query entry point as the query engine. The citizen profile would resume of what is happening in given locations. It would accept, for instance, location input and would reply with all information available concerning all topics for that location: urban and architectural plans, land expropriation, etc. The location selection could be via a point in a map, via item selection in list of locations, or via user typed locations.

The business profile would provide business opportunities in the municipal administrative area. Using the same knowledge base, a construction company could be more interested in knowing all open calls for constructors instead of knowing all information about a single construction. By default, the profile would show all open calls and other relevant information for businesses. It could also allow selecting locations like the citizen profile.

Finally, the generic query entry point would allow performing any query using natural language or some query language. This entry point could be used by citizens or businesses familiar with the system, by other parties that co-provide information and services, and by municipal staff that possess a good knowledge about the information system. The municipal staff can use this entry point as support for preparing internal reports. For instance, they could query the total amount of money spent in construction during a given year. The way a potential system would be deployed in such scenario is depicted in Figure 3.5.

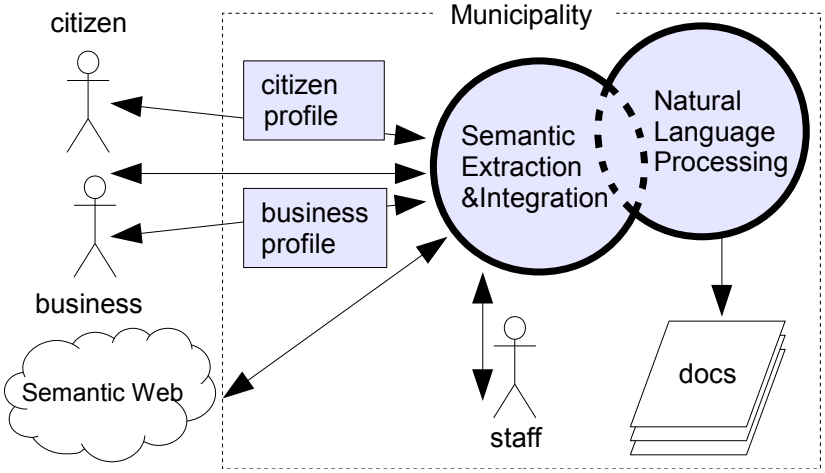


FIGURE 3.5: System managed by the information producer scenario: government services define the semantics and provide data examples. Information can be accessed via specific profiles or directly through a query entry point. Profiles use the query entry point as the query engine. The query entry point is part of the Semantic Extraction & Integration component.

3.4.2 Managed by an Information Consumer

This scenario describes how private institutions benefit from using the system to acquire information from government documents. The role of the system is to detect and acquire relevant information in government services websites that are related to the company business, in terms of opportunities in a given geographical area, such as municipalities, and/or legal information like legislation. The information is structured according to the semantic defined by the company and then provided to employees or shared with other enterprise information systems (see Figure 3.6). Examples of enterprises that could use this system are news corporations, construction companies, and suppliers of software, hardware, furniture, etc.

Company employees access information via natural language interface. Expert employees and enterprise information systems of the company access information using a query language entry point. Similarly to the previous scenario, the natural language interface uses the query entry point as the query engine.

Taking a closer look at local government level, businesses benefit from automatically tracking municipal information. It is possible for the system to track several municipalities simultaneously. A web crawler could periodically obtain all new published documents. The system would learn, from given examples, how each municipality outputs its data and thus could store information of different municipalities in the knowledge base. Triggers could be fired every time the system detects relevant topics as open contract calls or equipment acquisitions.

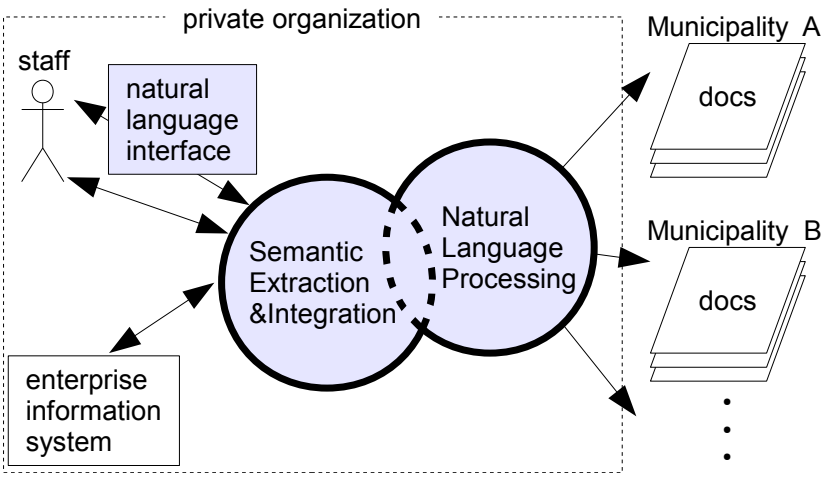


Figure 3.6: System managed by the information consumer scenario: private organizations define the semantics and provide data examples. They can track the information systems of several municipalities simultaneously. Employees access information via natural language interface or via a query entry point. Other enterprise information systems access information via the query entry point.

3.4.3 Managed by a Third Party

Besides being managed by information producers or consumers, the system is also useful in a scenario where it is managed by a third party organization. Associations concerned with citizenship, or nongovernmental organizations with specific concerns, can configure the system to track, collect, and integrate information relative to their areas of interest. As discussed in Section 3.1, third parties can play an important role in co-provisioning government services and information, in increasing data and infrastructure redundancy, and thus adding safeguards and reducing the dependency on government ICT infrastructure.

For instance, parents' associations can build a website with information fed by the system. The system can be configured to follow subjects like timetables of public sport facilities, parties and other events for children (e.g. when the circus is in town), news related to school, etc. The semantically aggregated information would then be permanently available to every household.

Similarly to the previous scenario, a web crawler could periodically obtain all new published documents. The system would learn how each selected institution outputs its data to be able to store it in the knowledge base. Citizens access information via a natural language interface or using a query language entry point.

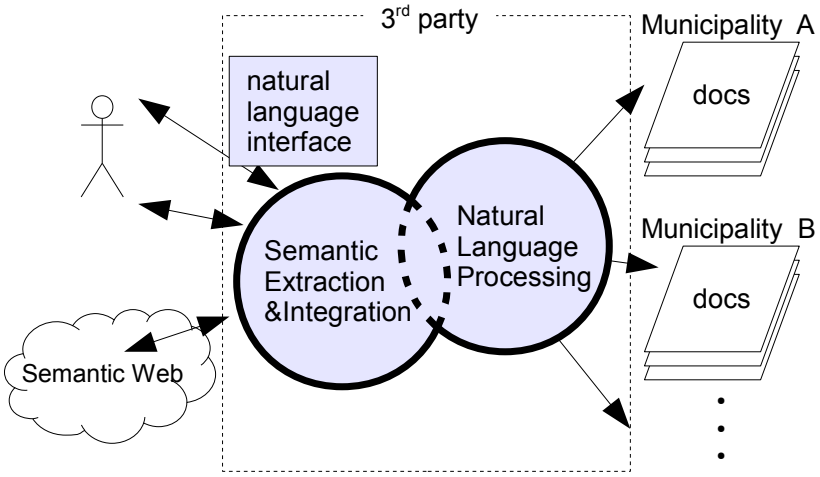


Figure 3.7: System managed by a third party scenario: the process of searching and integrating information is targeted for specific subjects relevant to a community. It is possible to track information systems of several municipalities simultaneously. Citizens access information via natural language interface or via a query entry point. Other enterprise information systems access information via the query entry point.

3.5 Summary

The proposed conceptual model addresses the five requirements that were identified based on the presented view of the future and on generic concerns regarding information provision.

The model is simple and clear, and outlines how an e-government information provision system should operate. It supports information acquisition from natural language documents, relevant information detection and integration, domain adaptation, user friendly interfaces, and data provision using open standards for interoperability purposes. It is a flexible platform that can be deployed in several scenarios and that can be used in a network of systems that co-provide information. The proposed model defines how different modules are arranged together and not how each module should operate internally. This option aims to allow future technology to be used without needing to change the model. It is possible to implement systems complying with this model using current technologies and standards.

The model acts as a hub that summarizes information gathered from different places in a convenient and unique place. The semantic knowledge can be queried and accessed instead or in complement of the original documents. It is fundamental to trace all information to the respective documents to make sure that citizens and businesses have all information they need to make informed decisions. The ability to change the knowledge domain increases the versatility of the system. Changes in laws and regulations can lead to changes in information provision. The modification can be in terms of the level of disclosure or the removal/addition of information types. It is important to be able to reflect, in an easy and timely manner, the changes in government needs without software re-engineering, saving time and money. However, this requirement imposes a technological option: the ability to learn from given examples. This implies having machine learning algorithms to identify data semantics.

The proposed model can be further extended. As it was designed for handling public documents and public information, one feature that was not included is user authentication. The lack of such feature prevents the usage of private information in systems strictly conforming to the model. Another concern in the model is the dependence on the ontology and semantic knowledge base. Current implementations of knowledge bases have performance issues when dealing with expressive ontologies and/or huge amounts of facts. However, current implementations are faster than earlier ones so it is expectable that the future brings technical solutions for this concern. Nevertheless, this doesn't prevent practical and useful systems to follow this conceptual model.

4

Resources and Tools

The present chapter describes the resources and software tools used in the development of a proof of concept prototype. The tools selected for language processing needed to process Portuguese language, or be adapted for it, as they will be used to build a prototype that needs to process documents written in Portuguese.

First are presented the resources used to adapt software tools for Portuguese and used to define the data semantics of the evaluation scenario. Then are presented the software tools that compose the proof of concept prototype and explained how the resources were used to train the tools. The explanation is in the respective tool subsection. The chapter ends with an overview of the chapter content.

4.1 Resources

Two types of resources were used in prototype development: (1) language resources used to prepare NLP tools for Portuguese which consist in an annotated corpus, a computational lexicon, and a database of lexical relations; (2) knowledge resources which are ontologies used for creating an ontology suitable for the prototype evaluation scenario.

Besides the resources described here, this work also used public documents acquired from public administration websites. They do not have a section in this chapter since they are regular documents that could be replaced by others without expected impact on achieved results. They were used to provide seed examples of semantic content and were the media from which the prototype acquired information. They are referred whenever relevant: while

describing the creation of seed examples and while evaluating the prototype.

4.1.1 Annotated Corpus

The annotated corpus used in this work was Bosque, a subset of Floresta Sintá(c)tica, a publicly available Treebank for Portuguese. Floresta Sintá(c)tica is a Treebank automatically built using the parser PALAVRAS and two daily newspapers corpora: newspaper Público from Portugal, and newspaper Folha de S. Paulo from Brazil (Afonso et al., 2002; Bick, 2000).

Bosque subset was fully revised by linguists and contains 9,368 sentences and about 186,000 words (Freitas et al., 2008). In this work was used Bosque v7.3 because it was the only Portuguese corpus found that was in a format suitable to train the syntactic parser, the Tenth Conference on Computational Natural Language Learning (CoNLL-X) format. CoNLL-X format defines a sentence as one or more tokens, each one starting in a new line and consisting of ten fields. Table 4.1 describes the fields by the same order they appear in each token, and Table 4.2 presents the first 20 lines of a training file. A comprehensive explanation of the POS tags and features can be found in Freitas and Afonso (2008).

Table 4.1: Description of CoNLL-X shared task token fields. Fields have the same order presented here.

Field name	Description
ID	Token counter starting at 1 for each sentence.
FORM	Word form or punctuation symbol.
LEMMA	Lemma or stem (depending on data set) of word form, or underscore if not available.
CPOSTAG	Coarse-grained part-of-speech tag. Tag set depends on the language.
POSTAG	Fine-grained part-of-speech tag. Identical to the CPOSTAG if not available.
FEATS	Unordered set of syntactic and/or morphological features (depending on the language), separated by a vertical bar (|), or underscore if not available.
HEAD	Head of the current token which is either a value of ID or zero ('0'). Depending on the original Treebank annotation there may be multiple tokens with an ID of zero.
DEPREL	Dependency relation to the HEAD. The set of dependency relations depends on the particular language. May be meaningful or simply 'ROOT'.
PHEAD	Projective head of current token which is either a value of ID, or zero ('0'), or an underscore if not available.
PDEPREL	Dependency relation to the PHEAD or an underscore if not available. The set of dependency relations depends on the particular language.

Bosque was made available by Linguateca, a distributed language resource center for Por-

Table 4.2: First 20 lines of Bosque v7.3 in CoNLL-X format.

1	Um	um	art	art	<arti>|M|S	2	>N	_	_
2	revivalismo	revivalismo	n	n	M|S	0	UTT	_	_
3	refrescante	refrescante	adj	adj	M|S	2	N<	_	_
1	O	o	art	art	<artd>|M|S	2	>N	_	_
2	7_e_Meio	7_e_Meio	prop	prop	M|S	3	SUBJ	_	_
3	é	ser	v	v-fin	PR|3S|IND	0	STA	_	_
4	um	um	art	art	<arti>|M|S	5	>N	_	_
5	ex-libris	ex-libris	n	n	M|P	3	SC	_	_
6	de	de	prp	prp	<sam->	5	N<	_	_
7	a	o	art	art	<-sam>|<artd>|S	8	>N	_	_
8	noite	noite	n	n	F|S	6	P<	_	_
9	algarvia	algarvio	adj	adj	F|S	8	N<	_	_
10	.	.	punc	punc	_	3	PUNC	_	_
1	É	ser	v	v-fin	PR|3S|IND	0	STA	_	_
2	uma	um	num	num	<card>|F|S	1	SC	_	_
3	de	de	prp	prp	<sam->	2	A<	_	_
4	as	o	art	art	<-sam>|<artd>|P	7	>N	_	_
5	mais	mais	adv	adv	<quant>	6	>A	_	_
...									

tuguese that aims to raise the quality of Portuguese language processing through the removal of difficulties faced by researchers and developers. This resource is difficult to create due to the time, human effort, and skills needed to build it with high quality. If this resource was not available, the tool selection relative to NLP would have to be different. As such, the existence of third party and public Portuguese language resources were an important contribute to this work.

4.1.2 Computational Lexicon

Bosque contains around 186,000 words and Floresta, the full corpus, contains around 300,000 words (Freitas et al., 2008). Tools like POS taggers benefit from broader lexicons as POS tagging algorithms have to guess tags of unknown words, which usually introduce some errors. To reduce the amount of tag guessing, and because the selected tagger allows to use separate files for corpus and lexicon, the lexicon was extended by using the computational lexicon LABEL-LEX-sw. LABEL-LEX-sw comprises more than 1,500,000 inflected word forms automatically generated from 120,000 lemmas and a set of generation rules (Ranchhod, 2005). Table 4.3 presents the first ten lines of LABEL-LEX-sw. Each line follows the format: `<inflected form>`, `<lemma>`. `<POS>` + `<Subcat>`: `<morphological attributes>`.

LABEL-LEX-sw uses a different tag set than Bosque and was converted to the tag set used by it. The conversion is based on LABEL-LEX-sw POS tags and, when necessary, subcategory information. Subcategories are used when a LABEL-LEX-sw POS tag is associated with two

Table 4.3: First 10 lines of LABEL-LEX-sw.

```
a, a. N + Letra + z1 : ms
a, a. PREP
a, ao. PREPXDET + Art + Def + z1 :fs
a, ao. PREPXPRO + Dem + z1 : fs
a, ei. Vmf : F4s : F3s
a, eu. PRO + Pes + z1 : L4fs : L3fs
a, o. DET + Art + Def +z1 : fs
a, o. PRO + Dem + z1 : fs
aa, a. N + Letra + z1 : mp
aacheniana, aacheniano. N : fs
...
```

or more Bosque POS tags. Table A.1 in Appendix A presents the outline of the conversion method. As it is straightforward to complete the information with LABEL-LEX-sw subcategories, for clarity purposes, the information is organized according LABEL-LEX-sw POS tags and subcategories were ignored.

4.1.3 Lexical Relations Database

Relations among lexical items reveal which items are similar or opposites which are part of others among others. Knowing such relations allows, for instance, organize natural language content and broaden the scope of information search. These type of resources exist for several languages of which a relevant example of such resource is WordNet (Miller et al., 1990).

In this work was used Palavras Associadas Porto Editora Linguateca (PAPEL), a lexical resource for Portuguese that contains semantic relations among lexical items such as hyperonymy, meronymy, and synonymy (Oliveira et al., 2008, 2009). PAPEL is a semantic network inspired by MindNet (Richardson et al., 1998). It was built by parsing a machine readable human dictionary, Dicionário da Língua Portuguesa, that is a commercial Portuguese dictionary owned by Porto Editora. Tables 4.4 and 4.5 present part of the relation set and the first ten lines of PAPEL, respectively.

PAPEL was selected over the WordNet for Portuguese, WordNet.pt (Marrafa et al., 2011), because it is a high quality resource and it is completely available. WordNet.pt database is not directly accessible and web services are not available. The only mechanism available is querying via website which was not considered suitable for integration in this project.

4.1.4 Ontologies

Four publicly available ontologies were used to create the ontology used in the prototype: (1) Friend of a Friend for identifying persons and its relationships; (2) Dublin Core to encode the references where the information was found; (3) Geo-Net-PT 01 to identify the political organization of places in Portugal complying with Geonames definition; (4) World Geodetic

Table 4.4: Sample of PAPEL relation set. Relations in boldface appear in the table which shows the first 10 lines of PAPEL, Table 4.5.

Type	Meaning	Example of relations
antonimia	antonymy	antonimo_adj_de , antonimo_n_de ... (2 more)
causa	cause	**accao_que_causa** , causador_de ... (3 more)
contido	contained, subset of	contido_em , contido_em_algo_com_propriedade.
estado	state, condition	**devido_a_estado** , tem_estado.
finalidade	purpose	faz_se_com , finalidade_da_accao ... (3 more)
hiperonimia	hyperonymy	hiperonimo_de.
local	place	local_origem_de local_onde.
maneira	manner, way of	maneira_de , accao_para_maneira ... (2 more)
maneira_sem	without manner	maneira_sem , maneira_sem_accao.
material	material, made of	material_de.
membro	membership	membro_de ... (2 more)
parte	part, meronymy	parte_de ... (2 more)
produtor	producer	produtor_de ... (2 more)
qualidade	quality, feature	tem_qualidade , devido_a_qualidade.
referente	referral, relative to	referente_a , diz_se_sobre diz_se_do_que.
sinonimia	synonymy	**sinonimo_n_de** , **sinonimo_v_de** ... (2 more)

Table 4.5: First 10 lines of the file with all relations of PAPELv3.2.

fatigar	SINONIMO_V_DE	desunhar
desinteligência	SINONIMO_N_DE	desunião
discórdia	SINONIMO_N_DE	desunião :: ;fig;
divisão	SINONIMO_N_DE	desunião
separação	SINONIMO_N_DE	desunião
desunido	DEVIDO_A_ESTADO	desunião
desunir	ACCAO_QUE_CAUSA	desunião
desarmonizado	SINONIMO_ADJ_DE	desunido
desunido	DIZ_SE_SOBRE	desacordo
separado	SINONIMO_ADJ_DE	desunido
...		

System 1984 for identifying latitude and longitude of locations. These ontologies were selected due to its pertinence for the evaluation scenario, and due to its widespread adoption and use.

Friend of a Friend

Friend of a Friend (FOAF) defines a variety of terms including people, groups, and documents (Brickley and Miller, 2010). In this work, FOAF was primarily used to describe people and thus most of the terms used are from the first of the three categories:

- Core - classes and properties describing characteristics of people and groups that are intrinsic to the subjects, such as classes `Agent` and `Person` and properties `name` and `knows`. In addition to these characteristics, FOAF defines classes for projects, organizations, and documents. Some of the classes and properties of FOAF correspond to other popular terms definitions, such as Dublin Core in the case of classes and properties relative to documents. Merging FOAF with other ontologies is straightforward as FOAF defines the correspondences using the standard Simple Knowledge Organization System (SKOS).
- Social Web - terms used to describe Web related concepts such as Internet accounts and address books.
- Linked Data - set of characteristics for information that is not readily summarized as simple factual data. These terms are pragmatical (e.g. `focus`, `LabelProperty`) and exist to support wider information linking efforts such as the Linked Data community.

Dublin Core Metadata Initiative

The Dublin Core Metadata Initiative (DCMI) provides the ability to describe resources such as documents and video (Weibel et al., 1998). In this work, Dublin Core was used to describe the documents were the information was found. The Dublin Core standard includes two levels and a group of element qualifiers that refine the semantics of the elements in ways that may be useful in resource discovery. The two levels of Dublin Core are: (1) Simple, comprising the 15 elements relative to content, intellectual property, and instatiation of resources (see Table 4.6); (2) Qualified, including the additional elements `Audience`, `Provenance` and `RightsHolder`.

Geo-Net-PT 01

Geo-Net-PT is a geographic ontology of Portugal (Lopez-Pellicer et al., 2009). In this work is used to obtain the organization of spaces, for instance which streets belong to a neighborhood, that in turn belongs to a city, that belongs to a municipality, etc.

Currently Geo-Net-PT is in its second version but, at the time of prototype development, the only version available was the first one and so the prototype uses Geo-Net-PT 01 (Chaves

Table 4.6: The 15 elements of Dublin Core Metadata Initiative.

Content	Intellectual Property	Instantiation
Title	Creator	Date
Subject	Publisher	Format
Description	Contributor	Identifier
Type	Rights	Language
Source		
Relation		
Coverage		

et al., 2005). The decision to maintain the first version was due to changes in the data model introduced by the second version. These changes would imply re-programming how Geo-Net-PT connects to the prototype without obvious benefits, as the first version serves well the role it has in the prototype. Table 4.7 presents the features that were relevant for this work.

Table 4.7: Features that exist in Geo-Net-PT. NUTS means Nomenclature of Territorial Units for Statistics. Source Lopez-Pellicer et al. (2009) and `http://linguateca.pt/geonetpt`.

Country	Province	Municipality	Settlement
Region	District	Zone	Street
Island	NUTS	Civil parish	Postal code

World Geodetic System

The World Geodetic System 1984 (WGS84) is a standard for use in cartography, geodesy, and navigation, and is the geodetic reference system used by Global Positioning System (GPS). In the prototype, WGS84 is used to store coordinates obtained from Google Maps. WGS84 has been updated since 1984 although the name has remained the same. It has two classes:

- `SpatialThing` - represents anything with spatial extent.
- `Point` - subclass of `SpatialThing` that describes points using a coordinate system relative to Earth.

These classes have three properties representing latitude (`lat`), longitude (`long`), and altitude (`alt`) information.

4.2 Tools

The used tools can be grouped in three categories: (1) NLP tools that assign formal structures to natural language plain texts; (2) knowledge engineering tools used to build a semantic representation and associate it to the formal structures assigned to natural language texts; (3) interface tools that make available the extracted semantic information for people and machines.

First are presented the NLP tools: sentence boundary detection, part-of-speech tagging, named entity recognition, and syntactic parsing. Next are described the knowledge engineering tools: ontology editor and semantic annotator. The section ends with the interface tools.

The performance of some tools is presented according to metrics frequently used in the evaluation of IE systems (Makhoul et al., 1999): precision, recall and F_1. Precision is the proportion of correct findings over all findings obtained, recall is the proportion of correct findings over all findings that existed in the documents, and F_1 is the weighted harmonic mean of precision and recall.

4.2.1 Sentence Boundary Detection

Sentence boundary detection addresses the problem of finding sentences boundaries. Sentences are fundamental and standard textual units in computational linguistics and, consequently, in natural language processing applications. Many linguistic phenomena such as collocations and variable binding are constrained by the abstract concept of "sentence" in that they are confined by its boundaries. However, finding these boundaries is not a trivial task since end-of-sentence punctuation marks are ambiguous in many languages including Portuguese. The period, which is employed often as sentence boundary marker, can also denote ordinal numbers, initials, abbreviations, or even abbreviations at the end of sentences. Exclamation points and question marks can occur within quotation or parenthesis as well as at the end of sentences (Palmer and Hearst, 1997; Kiss and Strunk, 2006).

Several approaches have been proposed over the years. The simplest implementation of rule based approaches consists in designing rules to find patterns that usually occur at the end of a sentence, such as "period-space-capital letter". Some works complement this approach with lists of the most common abbreviations. The machine learning proposals include several different techniques as decision trees, neural networks, and support vector machines. Table 4.8 presents a summary of some relevant works, the approach followed, and obtained performance regarding the corpora: the Wall Street Journal section of Tipster (T-W) (Harman and Liberman, 1993) and LacioWeb (Aluísio et al., 2003) for F_1 measure, and Brown corpus for American English (Francis and Kucera, 1979) and Wall Street Journal (WSJ) (Paul and Baker, 1992) for error rate. The performance results were obtained from the work of Kiss and Strunk (2006) and, as it can be seen, the F_1 measure for Portuguese – LacioWeb corpus – is above 0.95 for the machine learning approaches.

The system selected to be included in the proof of concept prototype was Punkt (Kiss and Strunk, 2006). Punkt's performance using other corpus is slightly inferior to the best results published in the literature. However, Punkt as the advantage of already being tested with Portuguese (Kiss and Strunk, 2006), and being provided with source code and detailed documentation.

Table 4.8: Summary of available sentence boundary detector software tools. Performance measured with the following corpora (Kiss and Strunk, 2006): Tipster Wall Street Journal (T-W), LacioWeb (Lacio), Brown corpus for American English, and Wall Street Journal (WSJ).

System	Approach	F_1		error	
		T-W	Lacio	Brown	WSJ
Grefenstette and Tapanainen (1994)	Rule based. Regular expressions + common abbreviation list.	–	–	.009	–
RE (Silla and Kaestner, 2004)	Rule based. Regular expressions.	.918	.899	–	–
Riley (1989)	Machine learning. Decision trees based on eight features.	–	–	.002	–
Satz (Palmer and Hearst, 1997)	Machine learning. Uses tags of tokens at left and at right of punctuation marks. Two learning methods tested: neural networks and decision trees.	.919	.992	–	.010
Reynar and Ratnaparkhi (1997)	Machine learning. Maximum entropy modeling. Multi-lingual version uses abbreviation list induced from corpus.	.912	.965	.025	.020
Mikheev (2002)	Machine learning. Twenty rule patterns filled automatically from unlabeled texts.	–	–	.003	.005
Punkt (Kiss and Strunk, 2006)	**Machine learning. Log-likelihood ratio testing if a token depend on the preceding one.**	**.915**	**.972**	**.010**	**.017**
Splitta (Gillick, 2009)	Machine learning. Eight features of tokens at left and at right of punctuation marks. Two learning methods tested: support vector machine and naive Bayes.	–	–	.007	.005

Punkt

Punkt uses a language independent, unsupervised machine learning approach and was tested with English and ten other languages including (Brazilian) Portuguese (Kiss and Strunk, 2006). It assumes that a large number of end-of-sentence ambiguities can be eliminated once abbreviations have been identified. It operates in two stages:

1. Produces an intermediate classification performed at the type level to detect abbreviation types and ordinary word types. This stage makes three assumptions: (1) abbreviations can be defined as collocations consisting of a truncated word and a final period, so it uses a modified version of Dunning (1994) encoding the assumption that the period usually occurs in the end of an abbreviation; (2) abbreviations are usually short. It makes the abbreviation candidates likelihood decline when their size increases; (3) abbrevia-

tions sometimes contain internal periods and thus likelihood of candidates with internal periods is increased.

2. Evaluates whether the results of the first stage have to be corrected or not. The decision is based on the immediate right context of every token with a final period. Three heuristics are employed in the decision process: (1) inspection of orthographic clues to detect sentence boundaries after abbreviations and ellipses. An example of orthographic clue is capitalization since a capitalized word usually indicates a preceding sentence boundary in mixed case text; (2) determination if two words surrounding a period form a collocation: if a period is surrounded by two words that form a collocation, the expectation is that it is not a sentence boundary marker; (3) inspection of words immediately after a period against a list of frequent sentence starters: if the word after the period is a frequent sentence starter, the expectation is that the period is a sentence boundary marker.

Although Punkt was already tested with Portuguese it was not possible to find a model ready to split Portuguese sentences. Thus, a model was trained using Punkt tools and a corpus with more than 6,500 sentences. Around half of the corpus sentences were randomly selected from Bosque, and the other half was randomly selected from municipalities' documents, the documents from which the prototype had to acquire information in the test scenario.

The trained model was briefly tested with sentences from Bosque and from municipalities' documents which were not included in the training corpus. The sentence splitting model was considered ready as the result obtained was $F_1 = 0.90$ which is inline with the ones reported in literature.

4.2.2 Part-of-Speech Tagging

Part-Of-Speech (POS) tagging is the task of labeling each word with its correct morphosyntactic category, its part of speech, such as noun, adjective, verb, preposition, etc. It is usually one of the earliest steps in NLP. There are two main difficulties in POS tagging. One is ambiguity because a significant amount of words have one or more POS. The other is the assignment of POS to words for which the tagger has no knowledge about. Typically POS tagging takes into account the context around the target word, within a sentence, and selects the most probable tag using information provided by the word and its context (Güngör, 2010).

Several approaches have been proposed over the years. Some publicly available implementations that have the potential to be used for tagging Portuguese language are presented in Table 4.9. The precision results were obtained in Branco and Silva (2004), Aires (2000), and in the LacioWeb project website[1]. These results were obtained using the corpora Bosque 8.0

[1] http://www.nilc.icmc.usp.br/lacioweb/ferramentas.htm

(Freitas et al., 2008), LacioWeb (Aluísio et al., 2003), LX-Corpus, Tycho Brahe (Alves and Finger, 1999), and the English corpus Penn Treebank (Marcus et al., 1993).

Table 4.9: Precision achieved by state-of-the-art POS taggers with different corpora (Aires, 2000; Branco and Silva, 2004). Corpora used in evaluation are, in Portuguese, LX-Corpus (LX-C), Tycho Brahe (Tycho), LacioWeb (Lacio), and in English Penn Treebank WSJ (WSJ). Freeling was evaluated in Portuguese using Bosque 8.0 (marked with (B)) and TreeTagger was evaluated in English using Penn Treebank (WSJ and Brown corpus, marked with (P)).

System	Approach	Precision			
		Lacio	LX-C	Tycho	WSJ
TreeTagger (Schmid, 1994)	**Binary decision tree (n-grams).**	**.942**	—	**.885**	**.964**(P)
Brill (Brill, 1995)	Transformation-based error-driven.	.912	.971	.888	.965
MXPOST (Ratnaparkhi, 1996)	Maximum entropy model.	.955	.971	.897	.966
TnT (Brants, 2000)	Trigram hidden Markov model.	—	.969	—	.967
SVMTool (Giménez and Màrquez, 2004)	Support vector machine.	—	—	—	.972
FreeLing (Padró et al., 2010)	Trigram hidden Markov model.		.955(B)		∼.97
Stanford (Manning, 2011)	Maximum entropy cyclic dependency network.	—	—	—	.973

Results do not show a clear lead, in terms of precision, of one system over the others. When considering the same corpus, precision differences between systems are less than 2 percentage points (0.02), except with LacioWeb where the gap is bigger. Also, some systems perform better than others with a corpus but worse with another corpus (see MXPOST and TnT for instance). All taggers were reported as adaptable to languages as Portuguese, and five of them have actually been tested with Portuguese. However, TreeTagger and FreeLing were the only systems for which a Portuguese language model was found. Moreover, TreeTagger also achieved good results with other languages, including Romance languages, which is an indicator of the robustness of its approach. These facts and the stable implementation of TreeTagger were the reasons for its selection.

TreeTagger

TreeTagger represents the probability of a tagged sequence of words as n-grams, which is given by the following recursive equation in the case of bi-grams and where w_x are words and

t_x tags:

$$p(w_1 w_2 ... w_n, t_1 t_2 ... t_n) = p(t_n|t_{n-2}t_{n-1})p(w_n|t_n) \times p(w_1 w_2 ... w_{n-1}, t_1 t_2 ... t_{n-1})$$

The difference of TreeTagger over other n-gram taggers is how transition probabilities are estimated ($p(t_n|t_{n-2}t_{n-1})$). It uses a binary decision tree encoding a n-gram model with less parameters to estimate than Hidden Markov Models. This implies that it needs less data to obtain reliable estimates of context transition probabilities avoiding sparse data problems (Schmid, 1994).

The decision tree automatically determines the appropriate size of the context - number of surrounding words - to be used in transition probabilities estimation. Possible contexts are n-grams and also more complex constructions such as: $tag_{-1} = ADJ \ \& \ tag_{-2} \neq ADJ \ \& \ tag_{-2} \neq DET$ (Schmid, 1994). The probability of the n-grams is given by the leaf that is reached when following the corresponding tree path starting at the root node (see Figure 4.1). The creation of the decision tree is performed using training corpus and two processing steps:

1. Construction - done recursively using an algorithm based on ID3-algorithm presented by Quinlan (1983). In each step is created a test dividing the samples in two subsets with maximal distinctness regarding the probability distribution of the predicted tag. The recursive expansion stops when the next test generates at least one subset of n-grams smaller than a given threshold.

2. Pruning - occurs after the construction step when both subnodes of a node are leaves and the information gain at the node is below some threshold. In this case the subnodes are removed and the node becomes a leaf.

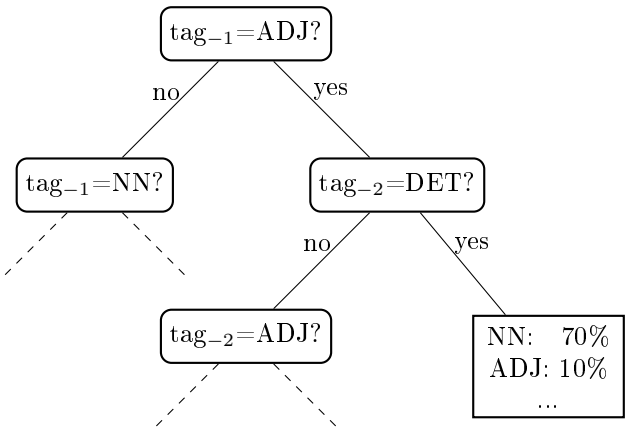


FIGURE 4.1: Example of a possible TreeTagger decision tree. Source Schmid (1994).

Training TreeTagger requires the creation of three files: (1) lexicon file, that was obtained from LABEL-LEX-sw by converting the tag set for the one used by Bosque (details in Section 4.1.2); (2) tagged training data, were used Bosque sentences with words and POS tags

as the only data fields; (3) open class file, a file containing the classes which the tagger can assign when guessing tags of unknown words. This file was kept the same as for English: `N ADJ V-FIN ADV`. Table 4.10 contains samples of each file, and Table 4.11 describes the tag set.

All parameters controlling the training process were kept with the default values: number of preceding words forming the tagging context (context length default is 2); threshold of information gain below which a leaf node of the decision tree is deleted (minimum decision tree gain default is 0.7); weight of the class of words with the same tag probabilities in the computation of the probability estimates (equivalence class weight default is 0.15).

Table 4.10: Samples of files used to train TreeTagger. Bosque and LABEL-LEX-sw formats were adapted to the ones presented here as corpus and lexicon respectively.

File	Sample of files content	
corpus	As	art
	associações	n
	servem	v-fin
	para	prep
	prestar	v-inf
	serviços	n
	a	prep
	a	art
	sociedade	n
lexicon	.	punc .
	,	punnf ,
	a	prep a art o
	associação	n associação
	associações	n associação
	naturais	adj natural n natural
open class	n adj v-fin adv	

Table 4.11 presents the set of tags used. The trained model was briefly tested with sentences from Floresta Sintá(c)tica not included in Bosque. The tagging model was considered ready as the results obtained, $precision = 0.92$, were inline with the ones reported in literature.

Table 4.11: Tag set used to train TreeTagger.

Tag	Part of Speech	Tag	Part of Speech
adj	adjective	pron-pers	personal pronoun
adv	adverb	prop	proper noun
art	article	prep	preposition
conj-c	coordinating conjunction	v-fin	finite verb
conj-s	subordinating conjunction	v-ger	gerund
in	interjection	v-inf	infinitive
n	noun	v-pcp	participle
num	numeral	punc	punctuation at end of sentences
pron-det	determiner pronoun	punnf	punctuation not at end of sentences
pron-indp	independent pronoun		

4.2.3 Named Entity Recognition

Named Entity Recognition (NER) seeks to locate and classify atomic elements in texts into predefined categories such as the names of persons, organizations, locations and so on. It involves processing a text and identifying certain occurrences of words or expressions as belonging to particular categories. The goal is identify several tokens like "New York City" as the single entity "New_York_City" which, in this case, represents a city. Identifying entities is a first step to understand who did what to whom, when and where. NER can be split into two sub-tasks: a segmentation task that finds the start and end of the sequence of words expressing an entity, and a classification task that labels the entity with a category according to the tag set used (Nadeau and Sekine, 2007; Bontcheva et al., 2009).

Rule based systems use hand-crafted rules to find word patterns corresponding to the named entities. The rules usually reflect the linguistic rules of the natural language being analyzed. Statistical based systems use large amounts of manually annotated data to train a statistical model that is later used to decide which groups of words form named entities. Early approaches to NER needed extensive gazetteers: lists of names of people, organizations, locations, and other named entities. The compilation of such gazetteers is considered a bottleneck in the design of NER systems because it requires a big effort to create and update such lists, so modern approaches avoid using gazetteers. State-of-the-art systems are performing NER over the web contents (Whitelaw et al., 2008).

Several NER systems were developed for Portuguese. A great part of them participated in the second, and so far last, NER evaluation contest for Portuguese named HAREM, organized by Linguateca (Mota and Santos, 2008). A summary of the participant systems, their approach and results achieved is presented in Table 4.12. Some systems are specialized in recognizing specific types of named entities, like geographic entities (Cage2 and SEIGeo) or temporal expressions (PorTexTO), and so their results were not as good for generic named entities recognition. Results show that REMBRANDT was one of the top performing systems in the task of generic NER. This, and the fact of being an open source software, made REMBRANDT the system selected for integration in the proof of concept prototype.

REMBRANDT

REMBRANDT tries to identify and classify each named entity according to the second HAREM directives (see Table 4.13). Alongside categories and types, it also assigns subtypes that are not presented because the proof of concept prototype does not use them. It operates in three stages using sets of rules and Wikipedia categories:

1. Identification - application of a first set of rules identifying numerical expressions, either numerical or textual. Then another set of rules identifies temporal expressions and values, based on the previously identified numerical expressions. This step ends with

Table 4.12: Summary of named entity recognition systems working for Portuguese. Precision (Prec.), recall (Rec.), and f-measure (F_1) values presented can be found in Mota and Santos (2008). Mota and Santos (2008) contains chapters describing each system.

System	Approach	Prec.	Rec.	F_1
Priberam	Ontology representing a lexicon with morphosyntactic and semantic classification + contextual rules.	.642	.515	.571
REMBRANDT	**Rules to identify named entities + Wikipedia structure to classify.**	**.650**	**.504**	**.567**
XIP-L2F/Xerox	Rule based dependency parser + rules to classify entities.	.657	.465	.545
REMMA	Rules and gazetteers + semantic categories of Wikipedia to classify candidates.	.605	.362	.453
R3M	Heuristics to identify and eliminate candidates. Machine learning to classify candidates.	.764	.252	.379
SeRELeP	Rules to identify entities and the relations between them.	.818	.242	.373
Cage2	Gazetteers for named entity and geographical identification + rules and regular expressions.	.450	.276	.342
SEIGeo	Regular expressions to detect geographical entities + geographical ontologies as knowledge sources.	.749	.117	.202
PorTexTO	Regular expressions to detect temporal expressions in sentences that contain at least one keyword of a list.	.679	.089	.156

the generation of more named entity candidates which are all sequences of tokens that contain, at least, a capital letter and/or an numeric digit.

2. Classification - starts by finding the Wikipedia page which title is more similar to the candidate text and collects the categories associated. Those categories are mapped to HAREM categories using a set of rules. Then, grammar rules based on textual clues from the candidates and surrounding tokens, supervise and possibly change the assigned HAREM category.

3. Second round of classification - relations between detected named entities are found using another set of rules. Those rules are used to classify not yet categorized named entities if they are related to named entities with category. In the end, all named entities without a category are discarded.

Table 4.13: Classes and types used in the second HAREM.

Categories	Types
abstraccao (abstraction)	disciplina (discipline); estado (state); ideia (idea); nome (name); outro (other)
acontecimento (occurrence)	efemeride (unique occurrence, news); evento (event); organizado (organized); outro (other)
coisa (thing)	classe (class); membroclasse (member of class); objecto (object); substancia (substance); outro (other)
local (place)	fisico (physical); humano (human, political); virtual (virtual); outro (other)
numero (number)	numeral (numeral); ordinal (ordinal); textual (textual)
obra (work)	arte (art); plano (plan); reproduzida (reproduced); outro (other)
organizacao (organization)	administracao (administration); empresa (enterprise); instituicao (institution); outro (other)
pessoa (person)	cargo (job, position); grupocargo (position category); grupoind (undefined group); grupomembro (group); individual (individual); membro (member of group); povo (people); outro (other)
tempo (time)	duracao (duration); frequencia (frequency); generico (generic); tempo_calend (calendar time); outro (other)
valor (value)	classificacao (classification, ranking); moeda (currency); quantidade (amount); outro (other)
outro (other)	–

4.2.4 Syntactic Parsing

Syntactic parsing is the task of analyzing a text made of a sequence of tokens to determine its structure with respect to a given grammar. One difficulty in syntactic parsing derives from the fact that natural language sentences often have ambiguous structure and meaning. This ambiguity is related to the high expressiveness of natural languages, and finding the most correct structure and meaning depends on the sentence and also on the context it appears (Section 2.2.2 elaborates on this topic). The grammars to syntactically parse natural languages can be defined with different formalisms that usually reflect both linguistic and computational concerns (Sag and Wasow, 1999; Jurafsky and Martin, 2008).

The grammar formalism found more adequate for the proof of concept prototype was dependency grammar. This decision was supported by two desired characteristics:

- Dependency grammars explicitly encode predicate-argument structures, which is useful in tasks such as information extraction.
- This approach led to the development of accurate syntactic parsers for several languages, particularly in combination with machine learning from syntactically annotated corpus. Such corpus exists for Portuguese.

Dependency grammars do not have phrasal nodes. Their structures are composed by words linked by binary, asymmetrical relations called dependency relations. These relations can be off different types and are hold between a syntactically subordinate word, called the dependent, and another word on which it depends, called the head (Kübler et al., 2009). Figure 4.2 depicts the dependency structure for a sentence. The arrows represent relations and are pointing from the head to the dependent. Arrows' labels encode the type of dependency relation.

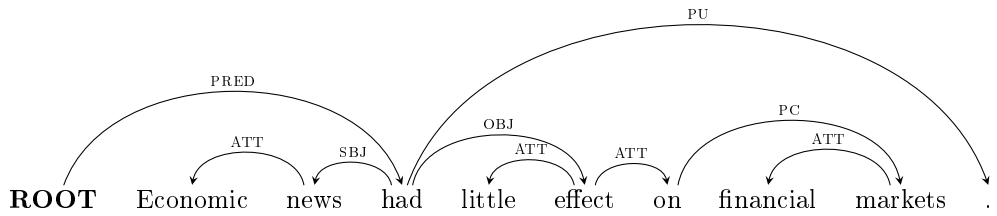


Figure 4.2: Example of a sentence dependency structure. Source Kübler et al. (2009, p. 2).

Multilingual dependency parsing was the topic of the shared task held by the CoNLL-X. The parsing systems were required to learn from training data, to generalize to unseen test data, and to handle multiple languages. Training sets were available for 13 languages, including Portuguese (Buchholz and Marsi, 2006). In the following year was added domain adaptation to the task (Nivre et al., 2007) but, as Portuguese was not included in the language set, results are not presented here.

Table 4.14 summarizes the CoNLL-X task results, presenting the results for Portuguese and for the average of the 13 languages (Màrquez and Klein, 2006). The system identification is the family name of the first author. The first column contains the values of Labeled Attachment Score (LAS), the main measure providing the proportion of tokens that are assigned both the correct head and the correct dependency relation label. The second column contains the values of Unlabeled Attachement Score (UAS) which is the proportion of tokens that are assigned the correct head, regardless of the dependency relation label. The third column contains the Label Accuracy (LA) which is the proportion of correct labels, regardless of the dependency relation head. Punctuation tokens were excluded from scoring.

TABLE 4.14: Results of the CoNLL-X shared task. LAS - proportion of tokens with correct head and dependency relation label. UAS - proportion of tokens with correct head regardless of dependency relation label. LA - proportion of correct labels, regardless of the dependency relation head. ([1]) no system description was provided.

System	Score for Portuguese (13 languages average)					
	LAS		UAS		LA	
Attardi	75.36	(61.23)	85.03	(76.17)	80.79	(70.72)
Bick	78.18	(70.00)	84.29	(77.52)	83.65	(80.27)
Canisius	77.42	(70.80)	85.61	(78.41)	81.85	(78.62)
Carreras	83.37	(74.72)	87.76	(81.19)	88.74	(83.53)
Chang	3.99	(76.80)	88.60	(83.54)	88.84	(84.10)
Cheng	85.07	(77.70)	90.30	(84.60)	88.00	(84.14)
Corston-Oliver	84.59	(76.94)	88.96	(84.37)	88.88	(83.96)
Dreyer	75.28	(65.23)	82.41	(74.47)	82.41	(75.22)
Johansson	84.57	(74.93)	88.40	(80.39)	89.42	(83.72)
Liu	71.13	(63.29)	77.10	(70.72)	76.46	(73.59)
McDonald	86.82	(80.27)	91.36	(86.61)	90.46	(86.65)
Nivre	**87.60**	**(80.19)**	**91.22**	**(85.48)**	**91.54**	**(86.75)**
Riedel	84.43	(77.94)	89.42	(85.04)	88.54	(84.85)
Schiehlen	71.01	(63.29)	81.27	(72.12)	77.14	(75.72)
Wu	81.47	(71.71)	85.57	(78.38)	87.16	(79.08)
Yuret	70.35	(65.01)	79.46	(73.47)	73.49	(70.87)
O'Neil ([1])	84.69	(78.43)	89.70	(85.30)	88.70	(85.05)
Sagae ([1])	86.01	(77.84)	89.78	(83.71)	90.22	(85.62)

The syntactic parsing is done with a data-driven dependency parser named MaltParser, that in Table 4.14 is identified as Nivre (Hall et al., 2007). MaltParser was selected because it was a top performing system in CoNLL-X shared task for all languages including Portuguese, and it is open source.

MaltParser

MaltParser is a system for data-driven dependency parsing, which can be used to induce a parsing model from Treebank data and to parse new data using that model. The system is based on four components (Nivre et al., 2006):

- Parsing algorithm – it builds a labeled dependency graph in one left-to-right pass over the input, using a stack to store partially processed tokens and adding arcs using four elementary actions: (1) `shift`, push the next token onto the stack; (2) `reduce`, pop the stack; (3) `right-arc(X)`, add an arc labeled `X` from the top of the stack to the next token, and push that next token onto the stack; (4) `left-arc(X)`, add an arc labeled `X` from the next token to the top of the stack, and pop the stack.
- History-based feature models – uses features of the derivation history to predict the next parser action. The features used are all extracted from the fields of the CoNLL-X data format, except from the fields ID and HEAD. The complete list of features is presented in Table 4.1.
- Map history to parser actions – in the setup used in CoNLL-X task, and in the proof of concept prototype, support vector machines were trained to predict the next parser action from a feature vector representing the history.
- Pseudo-projective parsing – deals with non-projective structures in a projective data-driven parser. Non-projective arcs are lifted one step at a time, and their label is encoded using a scheme that allows recovering non-projective dependencies, if necessary, by applying an inverse transformation to the output.

The parsing algorithm used was the same of the Single Malt system, a pseudo-projective dependency parsing with support vector machines which perform in linear time, and use a stack to store partially processed tokens and an input queue of remaining tokens (Hall et al., 2007; Nivre et al., 2006). The parsing model for Portuguese was induced with Bosque v7.3 used in the CoNLL-X shared task: multi-lingual dependency parsing.

4.2.5 Ontology Creation and Edition

Ontology editors are tools that provide assistance in the process of creation, manipulation, and maintenance of ontologies. They can work with various representation formats and, among other things, ontology editors provide ways to merge, visualize, and check the semantic consistence of ontologies.

Despite ontologies can be created using any text editor, having efficient tools speeds up processes and allows developers to focus in what is essential in applications. Ontologies can be difficult to build because they are formal models of human domain knowledge that is often tacit, and not rarely there are more than one possible mapping of that knowledge into formal, discrete structures. Although some rules of thumb exist to help in ontology design, it is more productive to have tools that, at least, identify simple conflicts and allow rapid re-design of ontology parts (Knublauch et al., 2004).

Several editors have being developed over the years and a summary of some available editors is presented in Table 4.15. In this work the selected ontology editor was Protégé

because it works stand-alone, and is an open source tool for which several plugins have being developed over time.

Protégé

Protégé is an open-source tool developed at Stanford Medical Informatics. Relevant features are the ability to assist users in ontology construction including importing and merging ontologies, the existence of several plugins that include alternative visualization mechanisms and alternative inference engines.

It is domain independent and its architecture clearly separates the internal representation of knowledge from its visualization. Protégé internal representation is comparable to object-oriented and frame-based systems. It represents ontologies consisting of classes, properties or slots, property characteristics such as facets and constraints, and instances. It features an open Java Application Programming Interface (API) to query and manipulate models and supports various data formats including RDF, OWL, and relational databases (Noy et al., 2000; Knublauch et al., 2004). Figure 4.3 is a screenshot of Protégé ontology visualization plugin OWL Viz.

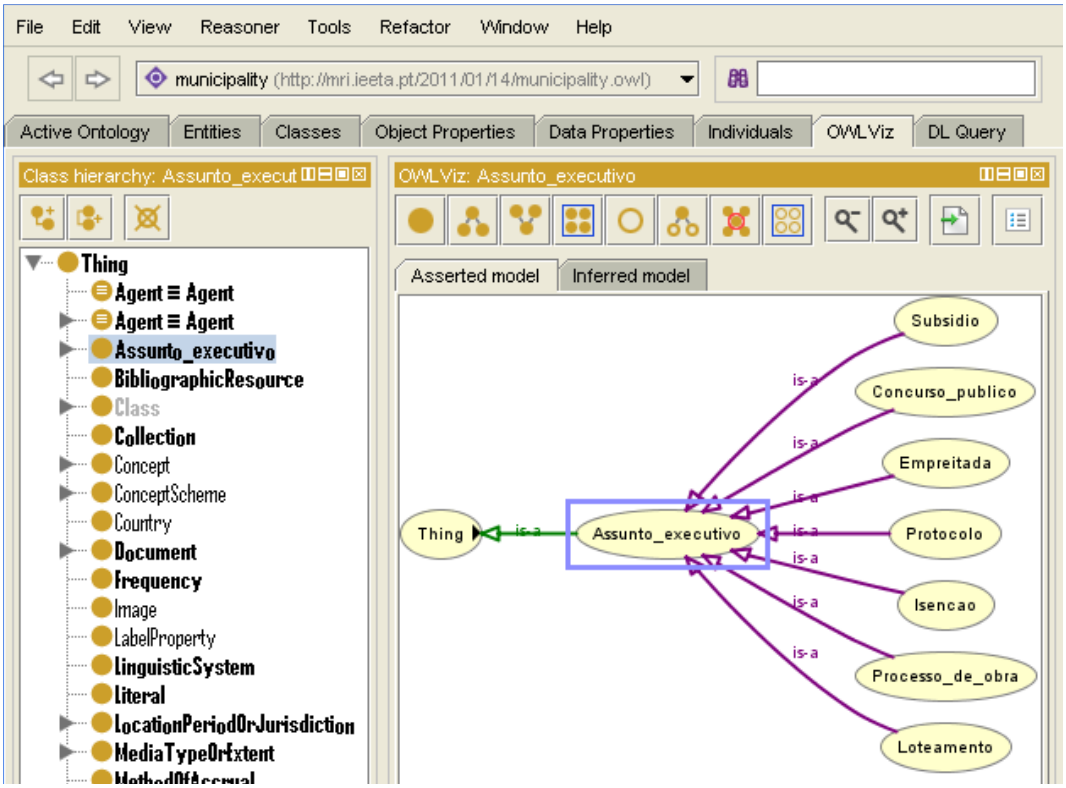


Figure 4.3: Screenshot presenting OWL Viz, a Protégé ontology visualization plugin.

Table 4.15: Properties of some ontology editors listed in World Wide Web Consortium (W3C) website in December 2012 (http://www.w3.org/wiki/Ontology_editors).

Name	Properties	Reference
Anzo	RDF and OWL ontology editor within Excel, generates ontologies from Excel spreadsheets, generate an initial ontology based on spreadsheet data and structure	http://cambridgesemantics.com
Knoodl	Free web application/service that is an ontology editor, wiki, and ontology registry. Supports creation of communities where members can collaboratively import, create, discuss, document and publish ontologies. Supports SPARQL queries, community-oriented ontology and knowledge base editor	http://knoodl.com
Neologism	Web-based, open source, supports RDFS and a subset of OWL, built on Drupal, online vocabulary editor and publishing platform	http://neologism.deri.ie
NeOn	Eclipse-based, open source, OWL support, several import mechanisms, support for reuse and management of networked ontologies, visualization, especially suited for heavy-weight projects (e.g., multi-modular ontologies, multi-lingual, ontology integration, etc	Haase et al. (2008)
Protégé	**Java-based, downloadable, Supports OWL, open source, many sample ontologies, popular, pluggable**	Knublauch et al. (2004)
SWOOP	Java-based, downloadable, open source, OWL Ontology browser and editor from the University of Maryland, small and simple ontology editor	Kalyanpur et al. (2006)
TopBraid	Eclipse-based, full support for RDFS and OWL, built-in inference engine, SWRL editor and SPARQL queries, visualization, import of XML and UML, multipurpose Semantic Web editor	http://topquadrant.com

4.2.6 Semantic Annotation

Semantic Annotation is the creation of a specific sort of meta-data that provides references between entities appearing in resources and domain concepts modeled in ontologies. The original content is thus enriched with machine readable information that is related to structured knowledge about domains. The formal identification of concepts and the relations between those concepts in unstructured or semi-structured natural language documents allows to bridge the gap between natural languages and computational representations. This makes the documents understandable by people and machines (Popov et al., 2003; Uren et al., 2006; Sanchez et al., 2011).

The process of creating the annotations can be manual, semi-automatic, or fully automatic. Manual annotation tools allow users to add annotations to documents and share these with others. Semi-automatic tools make annotation suggestions based on previous annotations and some pre-defined rules, and then someone using the system decide which annotations are to be included. Fully automatic tools perform all annotations without any intervention. Automatic and semi-automatic annotation tools are usually ready to annotate English texts, and annotating Portuguese texts requires some adaptation. As most algorithms are based on text patterns and/or shallow syntactic analysis, as consequence, they are not able to annotate arbitrary relations between entities (Oren et al., 2006; Sanchez et al., 2011).

Table 4.16 summarizes the properties of some relevant semantic annotators. In this work it was selected the AKTiveMedia annotator, a semi-automatic annotator. As the conceptual model specifies that the knowledge domain is defined by people – a system administrator – this annotator will be used to create seed examples for training the semantic models. AKTiveMedia Annotator was selected because, among other things, it works with OWL ontologies, it has an intuitive GUI, and it is open source and thus could be modified to our specific needs.

AKTiveMedia annotator

AKTiveMedia is an open-source tool which supports annotation of text, images and HTML documents. It supports different types of annotations like ontology based annotations as well as free comments, and learns from previous annotations to make suggestions in order to ease the annotation task.

AKTiveMedia was designed having as a major concern its usage by people. As such, it provides a GUI that guides users through the annotation process reducing the complexity of the task. One relevant example is the ontology not being completely displayed by default. Its details are hided until they are necessary. For instance, properties are not displayed until users select a class. Then, only the properties which domain includes the selected class are displayed. Figure 4.4 presents a screenshot of AKTiveMedia annotation interface with a real document from the municipality of Águeda.

The human annotator starts by highlighting parts of text and assigning ontology classes

Table 4.16: Relevant semantic annotators. Annotator types are given by two letters, first for annotation type: Manual, Semi-automatic, Automatic; second for learning type: No learning, Unsupervised, Supervised.

Name	Type	Ontology format	Documents format	Reference
AKTiveMedia	**SS**	**RDF, OWL**	**HTML, images**	Chakravarthy et al. (2006)
COHSE	SN	DAML+OIL	HTML	Bechhofer and Goble (2001)
Melita	SS	RDFS, DAML+OIL	text	Ciravegna et al. (2002)
OntoMat	SS	DAML+OIL, OWL, SQL	HTML. Deep Web	Volz et al. (2004)
S-CREAM	SS	DAML+OIL	HTML	Handschuh et al. (2003)
Running SHOE	SN	SHOE	HTML	Heflin and Hendler (2001)
Annotea	MN	RDF	HTML, XML	Kahan et al. (2002)
CREAM	MN	DAML+OIL	HTML	Handschuh et al. (2003)
Mangrove	MN	RDF	HTML, email	McDowell et al. (2003)
Vannotea	MN	XML	Direct3D, jpeg2000, mpeg-2	Schroeter et al. (2003)
Armadillo	AU	RDFS	HTML	Ciravegna et al. (2004)
C-Pankow	AN	HTML	text	Cimiano et al. (2005)
MnM	AS	RDFS	HTML, text	Vargas-Vera et al. (2002)
KIM	AN	RDFS, OWL	HTML	Popov et al. (2003)
KnowItAll	AU	–	HTML	Etzioni et al. (2005)
SmartWeb	AU	RDFS, OWL	text	Buitelaar and Ramaka (2005)

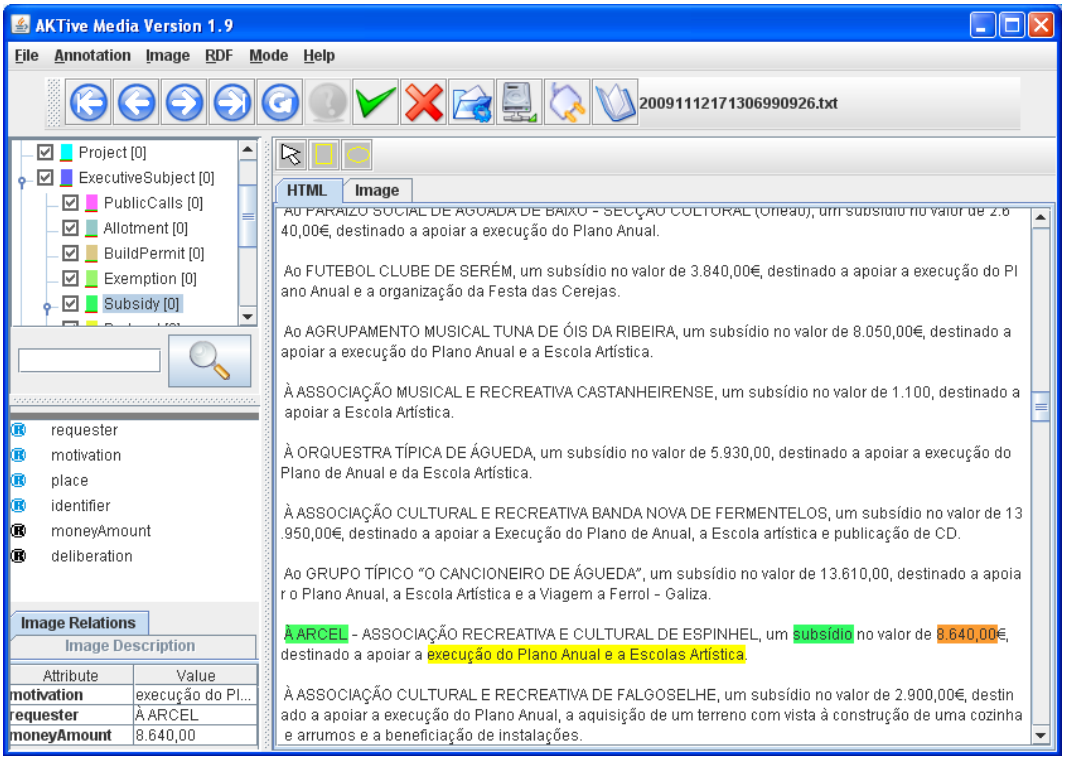


FIGURE 4.4: Screenshot presenting the semantic annotation interface of AKTiveMedia. The top left pane shows some ontology classes with the class `Subsidy` selected, below are visible the possible properties for the selected class, and the latest instances assigned to the class and its properties. The large pane at the right shows the document text displaying the annotated text highlighted in different colors which correspond to the respective ontology classes.

to the highlighted parts. Each part can become the subject of an ontology relation, which domain includes the class correspondent to the selected text. For that, it is necessary to select the highlighted text, select a relation in the relation panel, and select the text corresponding to the relation object.

4.2.7 Natural Language Interface

Natural language interfaces let users interact with computer systems using a familiar and intuitive way of communication. Section 2.2.3 provides a discussion about this topic. Here are presented the NLIs available to be used in the proof of concept prototype, and explained the choice that was made.

The performance of NLIs is measured in terms of the ability to map natural language inputs to the knowledge base structure - the recall - and how correct are those mappings - the precision. Some implementations require some customization to the usage domain before its usage, aiming to improve system performance, usually increasing recall. The performance of any NLI can depend of the usage domain (Damljanovic and Bontcheva, 2009; Kaufmann and

Bernstein, 2007).

Another important aspect of NLI in this work is natural language portability. Most implementations were developed for English and no tests were found in literature using Portuguese language. Porting between natural languages is a problem since NLIs usually have syntactic parsing and interfaces algorithms depend heavily on parsing results.

Table 4.17 summarizes the precision and recall values of some relevant NLIs, for different English datasets and without any domain customization. Unlike the others NLIs, NLP-Reduce and QuestIO do not perform syntactic analysis, which make these interfaces more easily portable to Portuguese. As it was not possible to find an implementation of QuestIO, NLP-Reduce was selected and tested for Portuguese, being integrated in the proof of concept prototype.

Table 4.17: Comparison of state-of-the-art natural language interfaces. Values for Precision and Recall without domain customization. Source Damljanovic and Bontcheva (2009).

Name	Mooney geography		Mooney restaurants		Software engineering		Geo. facts Germany		Reference
	P	R	P	R	P	R	P	R	
Aqualog	–	–	–	–	.864	.594	–	–	Lopez et al. (2007)
NLP-Reduce	.707	.764	.677	.696	–	–	–	–	Kaufmann et al. (2007)
ORAKEL	–	–	–	–	–	–	.842	.537	Cimiano et al. (2007)
Panto	.881	.859	.909	.966	–	–	–	–	Wang et al. (2007a)
Querix	.861	.871	–	–	–	–	–	–	Kaufmann et al. (2006)
QuestIO	–	–	–	–	.821	.891	–	–	Damljanovic et al. (2008)

NLP-Reduce

NLP-Reduce is a domain independent NLI that accepts full sentence queries, sentence fragments, or keywords in a text field. The query lexicon is automatically built from the knowledge base by extracting all explicit and inferred subject-property-object triples. The synonyms of all lexicon words are obtained from WordNet and are added to the lexicon (Kaufmann et al., 2007).

Porting NLP-Reduce to Portuguese involved the definition of stop words, wh-pronouns equivalents, and a small set of some specific keywords like sum and maximum. These definitions were explicitly coded in NLP-Reduce source code. The other change was PAPEL replacing WordNet in synonym search.

User inputs are first reduced by the removal of all stop words and punctuation marks. The remaining words are stemmed and passed to a query generator that will use them to produce a SPARQL query. The knowledge base upon which the interface will work is read at startup time, and the query production is a four step process (Kaufmann et al., 2007):

1. Search for triples that contain one or more words of the query in the object property label. Triples are ranked according to the amount of words included in the label.

2. Search for properties that can be joined with the triples found in step 1. Properties are searched using domain and range information of triples from step 1 and the remaining query words. In the case of query words producing triples based on alternative object properties, the triples favored are those with highest score from step 1. The triple set of this step is combined with the set of step 1, conforming the ontology rules.

3. Search for datatype property values that match the query words not matched in steps 1 and 2. Found triples are once again ranked considering the amount of words included in the property values. All found triples respecting the domain and range restrictions of the set created in step 2 are added to it.

4. When there are no more query words left, it is generated the SPARQL query for the join of the retrieved triples that achieved the highest scores in steps 1 to 3. Semantically equivalent duplicates are removed and the query is ready to be passed to a SPARQL endpoint.

4.3 Summary

This chapter presented the resources and tools used to develop the prototype. The resources, namely the annotated corpus, were chosen considering restrictions relative to the language, Portuguese, and to the tool selection. The availability of the corpus allowed the prototype to benefit from high quality software that was not originally developed to handle Portuguese. The choice of ontologies privileged well known resources to facilitate the semantic interoperability of the prototype. This selection guarantees that, at least, a part of the knowledge base is readily understood by other reasoning software, and also by other software developers, researchers, and practitioners.

The main processing tools for the development of the proof of concept prototype were introduced. The tools cover different research areas and were considered the most appropriate for this prototype at the time of decision. This implies that the tool selection can be reviewed as the state of the art in those research areas advances.

5

Prototype Implementation

The present chapter describes the implementation of the proof of concept prototype. Section 5.1 provides an overview of the prototype architecture. Section 5.2 illustrates how natural language processing tools were integrated in the prototype. In section 5.3 is described how was built the domain representation used in the experiments, and in Section 5.4 is explained how semantic models are created and used. The chapter ends with a summary in Section 5.5 A description of how to setup an application using this prototype can be found in Appendix C.

5.1 Prototype Overview

The proof of concept prototype implementation follows the proposed conceptual model and is organized in three modules instantiating the three components of the conceptual model, respectively:

Natural Language Processing – includes document content processing technologies to retrieve structured information from natural language texts. It is named NLP because it is based on technologies from the NLP area. In this part of the prototype, text is extracted from documents and enriched with the inclusion of POS tags, identification of named entities, and the construction of syntactic structures.

Domain Representation – has tools for defining data semantics and associate it with samples of the NLP module output. System semantics is defined via ontology and, after the

ontology defined, is necessary to provide examples of ontological classes and relations in sample documents. The examples are used to train semantic extraction models.

Semantic Extraction and Integration – trains and applies semantic extraction models to all texts in order to obtain meaningful semantic information. It complements the extracted information with external structured sources, e.g. geocodes, and stores everything in a knowledge base conforming to the defined ontology. Information in the knowledge base can be obtained using natural language queries or via a SPARQL endpoint.

Figure 5.1 depicts the global architecture of the prototype. The following subsections will elaborate on each part of the prototype. The explanations are illustrated using the fragment of a minute of municipal council meeting presented in the first row of Table 5.1.

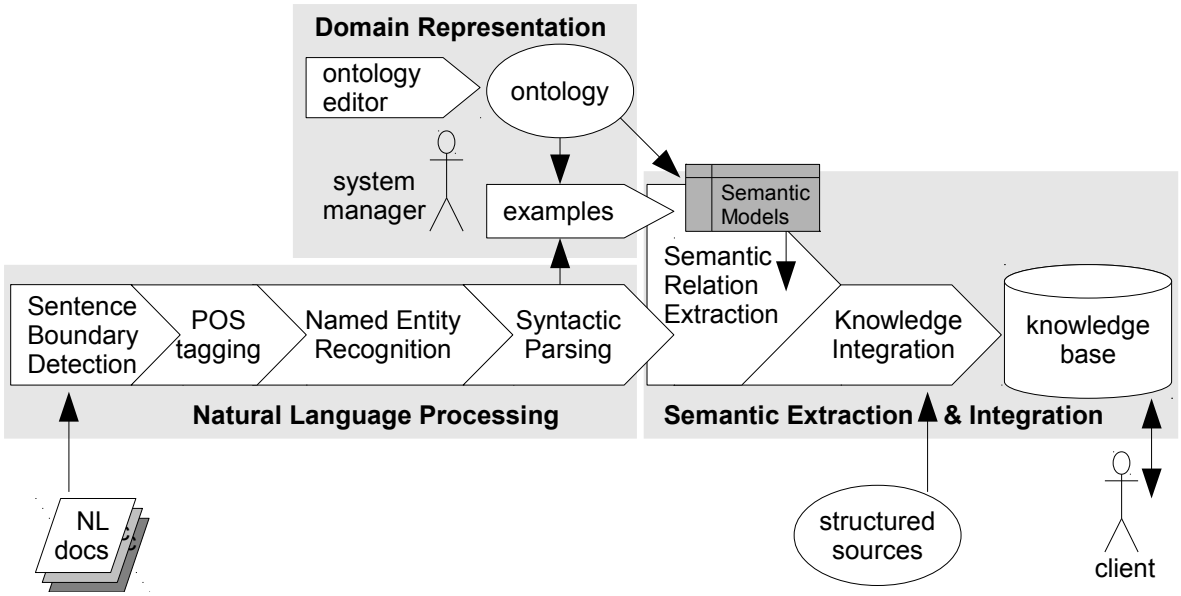


Figure 5.1: Proof-of-concept prototype architecture. The three parts of the conceptual model are delimited by the gray areas: NLP, Domain Representation (DR), and SEI.

5.2 Natural Language Processing

The prototype was developed having as background scenario the processing of documents written in Portuguese. Thus the NLP part was developed to handle Portuguese language. The processing is organized in four sequential steps: end of sentence detection, POS tagging, NER, and syntactic parsing.

Text is extracted from documents, and after sentences are separated using the sentence boundary detector Punkt. The sentence boundary detector step is relevant as all natural language processing will be done in a per sentence fashion. This means that sentences define

Table 5.1: Example of the output of the system modules.

Fragment:	"...atribuir os seguintes apoios financeiros:... À ARCEL - Associação Recreativa e Cultural de Espinhel, um subsídio no valor de 8.640,00€, destinado a apoiar a execução do Plano Anual e a Escola Artística"				
Translat.:	"... to award the following financial aid: ... To ARCEL - Associação Recreativa e Cultural de Espinhel, a subsidy amounting to €8,640.00, to support the implementation of the Annual Plan and the Art School".				
NLP out:	1	ÃĂ_Arcel	prop	0	UTT
	2	–	punc	1	PUNC
	3	Associação_Recreativa...	prop	1	N<PRED
	4	,	punc	1	PUNC
	5	um	art	6	>N
	6	subsídio	n	1	N<PRED
	7	em	prep	6	N<
	8	o	art	9	>N
	9	valor	n	1	N<PRED
	10	de	prep	9	N<
	11	8.640,00€	num	9	N<
AKTive M.:	example Subsidio:subsídio pretendente Organizacao:À_Arcel example Subsidio:subsídio montante montante:8.640,00				
KB entry:	<owl:NamedIndividual rdf:about="<URI>#s_8.64000eur"> <rdf:type rdf:resource="<URI>#Subsidio"/> <montante>8.64000eur</montante> <pretendente rdf:resource="<URI>#a_arcel"/> <terms:isReferencedBy rdf:resource="<URI>#acta_6902"/> </owl:NamedIndividual>				

the processing context in the next NLP steps, which include algorithms able to use all content of a sentence and that will not use any content of the previous or following sentences.

After split, sentences are enriched with POS tags assigned by TreeTagger: noun, verb, adjective, etc. TreeTagger was trained with a European Portuguese lexicon in order to be integrated in the system (details in Section 4.2.2). Its outputs have the format of one word per line, and each line contains the word form followed by the assigned POS tag and the word lemma.

Named entities are discovered and classified by REMBRANDT (Cardoso, 2008). Words belonging to a named entity are grouped using underscores. For instance, the names of the person *John Stewart Smith* become the single token *John_ Stewart_ Smith*. Unclassified named entities are collected to be classified using other strategies or discarded if other strategies fail. In the current implementation of the prototype, the alternative strategy is to query Google Maps for a location and, if a location is retrieved, the named entity is classified as an entity with a fixed physical location. Such entity can be an organization headquarters, a place that can be physical or human, or an event that happens always in the same place (details of NER classes in Table 4.13).

Outputs of POS tagging and NER are used to generate the syntactic parser input, which needs to be in CoNLL-X format. Inputs will have words of named entities as a single token since they represent a single concept. The generation of the syntactic parser input is done in seven steps (see Table 4.1 for details about the format):

1. Word forms (FORM field) of named entities are the named entity words group with underscores. Word forms of other words will be the words themselves.

2. Lemma of named entities will be its word form (words grouped with underscores). Lemma of other words will be the lemma given by TreeTagger.

3. Coarse grained POS tag (CPOSTAG field) for named entities is `prop` as, by definition, named entities are proper names. For the remaining words, this field is inferred from TreeTagger output as given by Table 5.2. This procedure is due to CoNLL-X need for two tags (CPOSTAG and POSTAG) when TreeTagger outputs one. Since POSTAG is a specialization of CPOSTAG, it is possible to infer CPOSTAG from POSTAG, but not the opposite. So, the approach followed was to assign TreeTagger POS tag to POSTAG, and to infer CPOSTAG from it.

4. Fine grained POS tag (POSTAG field) for named entities is `prop`. For all other words is the POS tag given by TreeTagger.

5. All punctuation tags (`punc` and `punnf`) become `punc`. TreeTagger needs a different tag for punctuation at the end of sentence, but the syntactic parser does not.

6. The ID field is the token position in the sentence. The ID of the first token of every sentence is 1, of the second is 2, etc.

7. All other CoNLL-X fields get dummy values represented by underscores as they will be filled by the parser.

Table 5.2: CPOSTAG field value of parser input assigned in function of TreeTagger output POS tag.

TreeTagger POS tag		CPOSTAG
conj-c; conj-s	→	conj
pron-det; pron-indp; pron-pers	→	pron
v-fin; v-ger; v-ing; v-pcp	→	v
any other tag	→	the tag itself

After merging the outputs of POS tagging and NER, sentences are analyzed to determine its grammatical structure. This is done by MaltParser (Hall et al., 2007). Figure 5.2 shows the dependency graph corresponding to the example output.

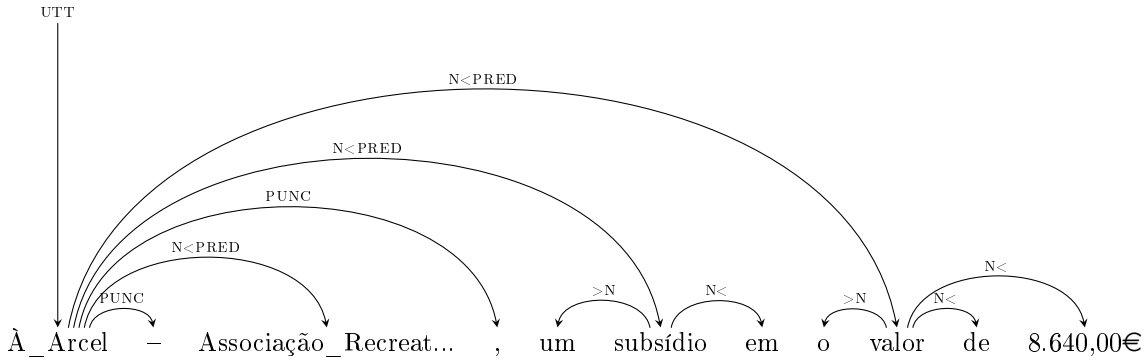


Figure 5.2: Sentence graph corresponding to the NLP output of the example.

The output after these steps is presented in the NLP output part of Table 5.1 (NLP out). In the example of Table 5.1, the word `um` was tagged as article (`art`), the word `subsídio` as noun (`n`) and so on and so forth. The first and third lines show that two named entities were found: `À_Arcel` and `Associação_Recreativa_e_Cultural_de_Espinhel`. The result of the third step of processing can be seen in the fourth and fifth columns of the NLP output. Fourth column denotes the dependency of the token (0 has no dependency; other number indicates the dependency of that token number) and the fifth column denotes the dependency type. For instance, the word `um` which is the token number 5 depends of the token number 6 which is the word `subsídio`. Moreover it is a dependency to a noun indicated by the symbols `>N` in the fifth column of token 5.

5.3 Domain Representation

The prototype semantics is formalized using an ontology. The ontology and the semantic extraction models associated to it define the domain representation of the prototype. Ontology is specified in OWL and, in the proof of concept prototype, is created using Protégé-OWL v4.1.

In this section, the definition of the ontology and its usage to provide seed examples to create semantic models is illustrated using examples obtained from the documents of the scenario defined to evaluate the prototype. The scenario is about municipal information delivery in which citizens query public information regarding their particular interest and regarding generic municipal information. The next chapter elaborates on the scenario and, in this chapter, is just relevant to know that citizens can ask several information regarding allotments, build permits, construction processes, and protocols.

The ontology defined combines the ontologies FOAF, Dublin Core, WGS84, and Geo-Net-PT 01. A new class named `ExecutiveSubject` was added to handle subjects relative to municipalities. Combining the four existing ontologies was done automatically in the following manner:

- Geo-Net-PT 01 did not share concepts with other ontologies and thus did not present merging challenges. Its classes and properties are just added to existing ones.

- FOAF have concepts related to Dublin Core and WGS84, although these last two do not share concepts between them. Since FOAF specifies its alignment with Dublin Core and WGS84 using the standard SKOS, Protégé was able to correctly align the three ontologies.

- The class `ExecutiveSubject` was defined as a subclass of the top level class `Thing`. It has seven subclasses or specializations, and the relations that their instances can have are defined by six object relations. (see Tables 5.3 and 5.4). These classes and properties were added to the ontology without merging issues just like Geo-Net-PT 01.

Disjunctions between classes and/or relations related to `ExecutiveSubject` were also defined. For instance, `ConstrProcess` is disjoint with `Protocol` but not with `BuildPermit` since both `ConstrProcess` and `BuildPermit` refer to construction and some overlap is possible.

Alongside with the ontology, it is necessary to provide examples of ontology classes and relations in texts. These examples are created using AKTive Media ontology based annotator (Chakravarthy et al., 2006) which is described in Section 4.2.6. Manual text annotation can become an error prone task as it can span over long periods of time, and/or requiring annotators searching for several different types of concepts simultaneously. To ease the manual annotation process, and thus helping reducing errors, the prototype features an automatic

Table 5.3: Subclasses of the class `ExecutiveSubject`.

Class name	Class description
Loteamento (allotment)	Permissions to perform land allotment or change previous allotments.
Empreitada (build permit)	Relative to construction contracts, public or private, already in execution.
ProcessoDeObra (construction process)	Announces relative to generic public construction processes: beginning of works, changes in budgets, expropriations, etc..
Isencao (exemption)	Requests for municipal fees exemptions.
Protocolo (protocol)	Protocols signed with institutions like schools or local clubs.
ConcursoPublico (public calls)	Announcements of public contracts to acquire equipment or build some facility.
Subsidio (subsidy)	Subsidies or allowances granted or requested.

Table 5.4: Relations which domain is exclusively `ExecutiveSubject`.

Relation name	Relation description
deliberacao (deliberation)	The outcome.
identificador (identifier)	The unique identifier given by services.
montante (money amount)	Any money amount involved in the process.
motivacao (motivation)	The motive.
local (place)	The address of the: construction/allotment, institution that signed the protocol, or entity that requests an exemption or subsidy.
pretendente (requester)	The entity or entities, excluding the municipality, that is/are involved in the process.

pre-annotator. Its usage is optional and its purpose is to highlight key text expressions before the annotation process starts.

Pre-annotations become effective annotations only after validated by a human annotator. Otherwise will be ignored in future processing stages. A configuration file containing regular expressions defines which text expressions are to be pre-annotated. The pre-annotation color scheme is the same of the annotation, being related to the ontology class. However, pre-annotations are not properly aligned with words in order to become distinguishable. Table 5.5 presents a possible configuration file.

Table 5.5: Example of a pre-annotation configuration file. Line format is: <regular expression> <tab character> <OWL class or classes separated by comma>.

[Cc]oncurso	ConcursoPublico
[Ee]mpreitada	Empreitada
[Ii]sen	Isencao
[Ll]ote	Loteamento
[Pp]rocesso(s){0,1}_(DE|de)_[Oo]bra(s){0,1}	ProcessoDeObra
[Pp]rocesso(s){0,1} (DE|de) [Oo]bra(s){0,1}	ProcessoDeObra
[Pp]rotocolo	Protocolo
[Ss]ubs.di	Subsidio

After the annotation process is completed is generated a text file containing a semantic relation per line. Each line starts with the keyword example followed by the relation subject coded as `subject_class: subject_text`. Then follows the relation name and the object of the relation coded the same way the subject of the relation (see fourth row of Table 5.1).

Individuals not involved in relations are ignored. As manual annotations should involve individuals (members of classes) with one or more properties (relations), individuals without properties were likely generated in a pre-annotation step, and thus should be ignored since they were not validated by a person.

Annotations involving more than one relation are broken into several relations. For instance, according to our ontology the sentence "... to award the following financial aid: ... To ARCEL - Associação Recreativa e Cultural de Espinhel, a subsidy amounting to €8,640.00, to support the implementation of the Annual Plan and the Art School" is a subsidy and contains four relations: `pretendente(subsídio, ARCEL)`; `montante(subsídio, 8640)`; `motivo(subsídio, Annual Plan)`; and `motivo(subsídio, Art School)`.

5.4 Semantic Extraction and Integration

The prototype creates one semantic extraction model for each ontology class and one semantic model for each ontology relation. This way a model represents a specific ontology class or relation. A model is a set of syntactic structure examples and counterexamples that were found to encode the meaning represented by the model. It also contains a statistical

classifier that measures the similarity between a given structure and the model internal examples. The model is said to have positively evaluated a sentence fragment if the similarity is higher than a given threshold. In runtime, all models evaluate all sentences and the fragments positively evaluated by a model are assigned the ontology class or relation represented by that model (Rodrigues et al., 2011a,b).

Unlike the previous two parts of the prototype, which have single usage sequences, this part have three different and interrelated ways of being used. The way it is used depends on the task to be performed: (1) creation of semantic extraction models; (2) using semantic extraction models to feed the knowledge base; (3) querying the knowledge base. The following subsections describe how these tasks were implemented.

Creation of Semantic Extraction Models

The algorithm for creating semantic extraction models was inspired in two works. The first work is about extracting instances of binary relations using deep syntactic analysis (Suchanek et al., 2006). In their study, Suchanek et al. (2006) extracted one-to-one and many-to-one relations such as place and date of birth. They have used custom built decision functions to detect facts for each relation, and a set of statistical classifiers to decide is new patterns are similar to the learned facts. In the proof-of-concept prototype, this work was extended to include the extraction of one-to-many and many-to-many relations. Also, as the conceptual model requires relations being defined by the ontology, the proof-of-concept prototype implements a general purpose decision function based on the annotated examples instead of a custom built function for each relation.

The second work is about improving entity and relation extraction when the process is learned from a small number of labeled examples, using linguistic information and ontological properties (Carlson et al., 2009). Improvements are done using class and relation hierarchy, information about disjunctions, and confidence scores of facts. This information is used to bootstrap more examples generating more data to train statistical classifiers. For instance, when the system is confident about a fact, as when it was annotated by a person this fact is used as an instance of the annotated class and/or relation. This fact can also be used as a counterexample of all classes/relations disjoint with the annotated class/relation, and as an instance of super-class/super-relation. Moreover, facts discovered by the system with high confidence score can be promoted to examples and be included in a new round of training. In the proof-of-concept prototype, this creation of more examples is not active by default as it can lead to data over fitting and should be used carefully.

Each semantic model contains a collection of partial syntactic structures. To obtain these structures, the file containing the manually annotated examples is read and the sentence that originated the examples are located and processed by the NLP part of the prototype. Then, each annotated example has the format `<subject class: subject text> <relation name>`

`<object class: object text>` and originates three facts:

1. `subject text` is an individual of class `subject class`;
2. `object text` is an individual of class `object class`;
3. `subject text` has relation `relation name` with `object text`.

Partial syntactic structures associated to the first two facts will be included in semantic models representing classes of the ontology, respectively the classes given by `subject class` and `object class`. To simplify, this type of model will be named entity-of models. Partial syntactic structures associated to the third fact will be included in the semantic model representing the relation of ontology that as the name `relation name`. This type of model will be named relation model.

Partial syntactic structures are represented by planar graphs that encode the dependencies given by the syntactic parser. The nodes of the graphs are the words of the sentence and the edges are the labeled dependencies (see Figure 5.3(a)).

Entity-of models associate subjects and/or objects to their ontological classes based on the connections between the subject/object and the other tokens of the sentence. The entity-of models store a collection of pairs for each subject/object. Two tokens are regarded as equivalent if they connect to the same lemmata using the same edges, although lemmata of nouns and adjectives are allowed to differ. Figure 5.3(b) depicts the three pairs stored by the entity-of model to characterize the token `subsídio`.

Relation models assess subject/object pairs based on the shortest graph path, called the bridge, between the elements of the pair. Two bridges are regarded as equivalent if they have the same sequence of nodes and edges, although nodes with nouns and adjectives are allowed to differ. Figure 5.3(c) depicts the bridge used by the relation models to associate `subsídio` and `8.640,00€`.

Semantic models also contain a statistical classifier that decides if previously unseen syntactic structures are similar to the ones stored by it. Structures considered similar enough are assigned the meaning of the model, otherwise are ignored. The statistical classifiers implemented in the prototype are based on the k-Nearest Neighbor (k-NN) algorithm, but others could be usedl (Rodrigues et al., 2011a,b).

The training process of the statistical classifiers belonging to the models starts by removing duplicate entries. Then counterexamples are searched. As it is assumed that all relations are marked in the sample documents, these documents are search for relation counterexamples. The search for counterexamples is not done for entities because entities (e.g. a person name found) can be an element of a class (class person) but is not involved in any relevant relation and thus is neither example nor counterexample.

Relation counterexamples are searched by having relation classifiers evaluating all sentences of the sample documents. After all sentences are evaluated, the positive evaluated fragments

that are not examples are considered counterexamples. Duplicate entries are removed and the remaining is added to the model. This process is repeated until the amount of counterexamples found is below a certain threshold level. The training procedure algorithm is further detailed in Appendix B.

Using Semantic Extraction Models to Feed the Knowledge Base

The procedure starts by loading all ontology triples. Triples have the format `<subject> <relation> <object>`, meaning that a given relation exists between the subject and the object. Following, all sentence graphs are evaluated by the classifiers of all semantic models, and are collected in the case of forming a triple. To consider that a sentence fragment forms a triple it is required to be positively evaluated by two entity-of models, one for subject and the other for object, and one relation model binding the subject and object.

Missing information according to the ontology is searched in external structured information sources. For instance, unknown locations of entities with a fixed place (as streets, organizations headquarters, and some events) are queried using Google Maps API. This procedure is somehow similar to the one used to disambiguate unclassified named entities but now is just used for individuals that exist in the knowledge base and that have a fixed location according to the ontology. Also, the political organization of the spaces - street $\subset$ neighborhood $\subset$ city $\subset$ municipality ... - is obtained using Geo-Net-PT01 (Chaves et al., 2005). This allows the system to display the information spatially on a map and to search and relate information by its location (Rodrigues et al., 2010c).

The information acquired from external structured sources was not obtained via semantic extraction models. This implies that forming triples from it involves writing specific code to transform that information in a valid triple for the ontology. It also implies that a change in the ontology probably implies a change in the custom built code. This prevents this way of acquiring information of having the same level of adaptability as the semantic extraction models, and thus should be used just when is strictly necessary.

All collected triples are temporary added to the knowledge base and their coherence is verified by a semantic reasoner. In the proof-of-concept prototype, reasoning is performed by an open source reasoner for OWL-DL named Pellet (Sirin and Parsia, 2004). All triples not coherent with the rest of the knowledge base are discarded, and a warning is issued. The remaining triples become part of the knowledge base.

The last row for Table 5.1 shows an entry of the knowledge base relative to the example. It is visible an `owl:NamedIndividual` of type `Subsidio` with some `montante`, with property `pretendente a_arcel` and `isReferencedBy acta_2009...` (in Portuguese "acta" means meeting minute). This entry defines a subsidy. Other entries not showed define the minute `acta_2009...` and the named entity `a_arcel`.

Querying the Knowledge Base

The knowledge base can be queried in two ways, one way is via a NLI that is more appropriate to humans, the other via a SPARQL endpoint that is more appropriate for machines. The NLI in the prototype is a re-implementation of the NLP-Reduce (Kaufmann et al., 2007) using Java Servlet technology (Oracle, 2013). The advantage of this re-implementation is to make NLP-Reduce available in a website.

The NLI creates a SPARQL query from the user input as described in Section 4.2.7. The SPARQL query is then passed to the SPARQL endpoint that is also available for third party systems querying the knowledge base. The output is then formatted in HTML to be displayed in the results page.

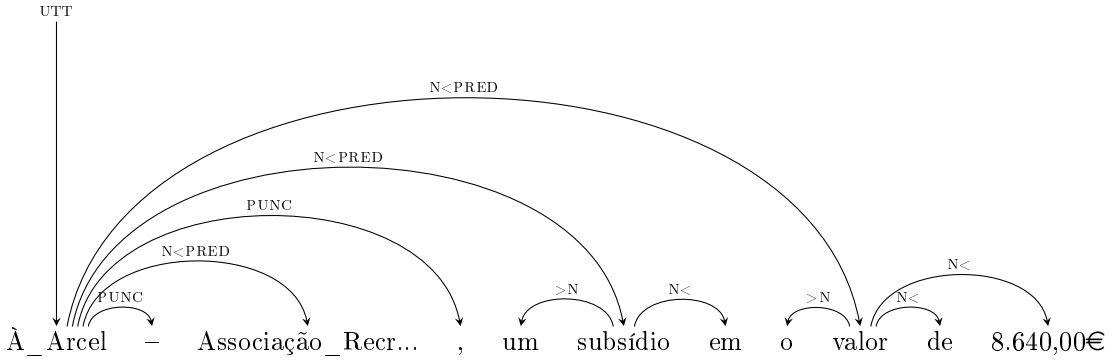
If the SPARQL endpoint is directly used, as by a third party machine, the output is in JSON format. The SPARQL endpoint used was the one provided with the Apache Jena triple store (Apache Jena, 2013).

5.5 Summary

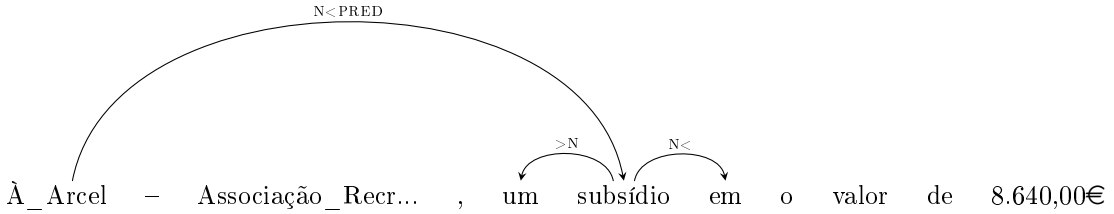
This chapter presented the architecture of the proof of concept prototype. The prototype follows the proposed conceptual model and its modular architecture can be seen as a framework where different modules can be plugged in and/or switched. This architecture provides flexibility to change the NLP pipeline in order to process other natural languages.

The prototype is able to assign semantic to information found in natural language documents. Moreover, the semantic is not defined *a priori*. System managers are able define the semantic, provide seed examples of relevant information and respective semantic, and the semantic extraction models algorithm learn how to extract similar information from the rest of the document set. This feature allows the prototype to be used in a wide range of scenarios such as local, regional and national governments.

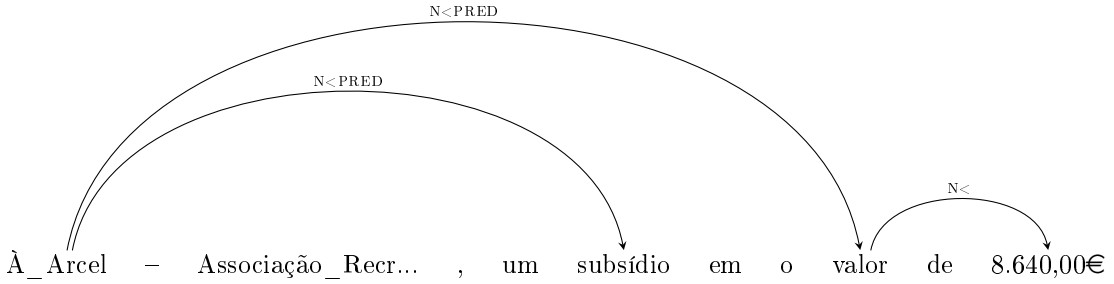
The prototype was built using high performance and open source software tools. Most of these tools were originally developed for other languages and, with the right language resources, were trained to work with Portuguese.



(a) The full sentence graph.



(b) The set of connections used by the entity-of model to classify `subsídio` as an element of the ontology class `Subsidio`.



(c) The bridge used by the relation model to associate `subsídio` and `8.640,00€`.

FIGURE 5.3: Graph of the syntactic structure of the fragment presented in Table 5.1.

6

Evaluation Results

This chapter reports the experiments conducted to evaluate the prototype operation and the results obtained in those experiments. The scenario outlined for tests fits in the scenario named "managed by a third party", being the third party represented by the author of this work (see Section 3.4.3 for details about the scenario). Here, the focus is on municipal information delivery as municipalities are often the closest point of service for citizens and enterprises. In this scenario, citizens query public information regarding their particular interests, and also query about generic municipal information. The decision about which topics were relevant was based on a meeting with municipality secretariat staff. The topics selected for citizens' particular interests were subsidies and build permits as these are the most requested and searched topics by individuals. Relative to generic municipal information, the topic selected was any kind of protocol because protocols can involve competency delegation, granting funds or other types of support, etc., becoming a mechanism that can have a deep impact on society.

The conceptual model followed by the prototype specifies two distinct and interconnected tasks: information provision with fully specified semantics, and information acquisition from natural language documents. As such, the prototype evaluation will encompass both of these tasks. The data used in testing each task is not the same because the software associated with these tasks was developed in different moments, and it was not changed after the tests reported here. The use of the same data for both tests would imply the use of non up-to-date information in the latest tests, or to repeat all tests with new data, which would consume time with no clear benefit. In fact, using documents of different municipalities and collected in different moments in time further stresses the prototype ability to work with different

document sources and, in this case, its independence of the municipality selected.

Regarding information provision, the conceptual model includes information access by people and by machines. It is assumed that machines have no problem accessing information since the knowledge base is accessible using public standards: ontology defined in OWL, and data access provided via SPARQL endpoint. In fact, all user interfaces developed for the current prototype use these methods to access the knowledge base with the same privileges granted to any third party machine. As for people access, the question is not if the information is there and if it's possible to obtain it. What is interesting to us is to know if the prototype brings some advantages over existing systems. This is a key aspect since systems as the prototype will only be useful if they improve over the current ones, the same is to say, if they are more usable than currently available systems. The usability experiment is the subject of Section 6.1.

Regarding information acquisition, the objective is to evaluate if the prototype is able to correctly detect and transpose relevant information to its knowledge base. These types of measurements are standard in the HLT research area. The objective is to assess if the prototype has a performance comparable to state of the art information extraction systems operating for the general domain. A difficulty when comparing with other systems is that no system able to extract entities and respective relations was found working for a generic domain in Portuguese. As such, is not possible to have direct comparisons and results presented always need to be interpreted having this restriction into consideration. This is the subject of Section 6.2.

Section 6.3 presents some concrete examples of possible applications for the proposed system, illustrating the potentialities of such systems and aiming to stimulate further developments. The chapter ends with a summary in Section 6.4.

6.1 Usability Assessment

An experiment was conducted to assess how potential users feel about using the proposed proof-of-concept prototype when searching for information. As observed by Barnum and Palmer (2011), if system designs fail to account for users' feelings about the system, users may be reluctant to adopt them. If users do not experience satisfaction, their attitude about using the system may affect the efficiency and effectiveness of their interaction with it.

One difficulty in finding how easy it is to use a system is that "...usability cannot be directly measured. Through operationalization of the usability construct, we find aspects of usability that can be measured." (Hornbæk, 2006). Usability is broadly considered a product of three elements: Effectiveness - can a task be well done and completed; Efficiency - did the resources available allowed finishing the task in a timely manner; Satisfaction - did users had a positive attitude towards the product (ISO 9241-11, 1998; Hornbæk, 2006; Travis, 2008). Effectiveness

and efficiency can be directly observable trough measurements of the system behavior given user inputs in conjunction with questionnaire (measure how much time a task took and query if participants accomplish their tasks, how many trial/wrong answers, etc.). Satisfaction is a more intangible element and thus more difficult to measure.

In this work, user satisfaction was assessed with Product Reaction Cards (Benedek and Miner, 2002). Product Reaction Cards were developed by Microsoft as part of a desirability toolkit. It is a set of 118 cards covering a wide variety of dimensions (see card list in Table 6.1). Each card contains a word or phrase, being around 60% of them positive and 40% negative or neutral. Participants in experiments using Product Reaction Cards are asked to pick the cards that best describe the product they have tested and how using the product made them feel. In our experiments the original cards were translated to the native language of participants, Portuguese, using the back-translation method. In this method two independent translators are involved: translator one translates the original version into the target language and then the second translator translates it back into the original language. Discrepancies were settled by consensus and the result is presented in Table D.1, Appendix D.

Table 6.1: The complete set of product reaction cards.

Accessible	Creative	Fast	Meaningful	Slow
Advanced	Customizable	Flexible	Motivating	Sophisticated
Annoying	Cutting edge	Fragile	Not secure	Stable
Appealing	Dated	Fresh	Not valuable	Sterile
Approachable	Desirable	Friendly	Novel	Stimulating
Attractive	Difficult	Frustrating	Old	Straightforward
Boring	Disconnected	Fun	Optimistic	Stressful
Business-like	Disruptive	Gets in the way	Ordinary	Time-consuming
Busy	Distracting	Hard to use	Organized	Time-saving
Calm	Dull	Helpful	Overbearing	Too technical
Clean	Easy to use	High quality	Overwhelming	Trustworthy
Clear	Effective	Impersonal	Patronizing	Unapproachable
Collaborative	Efficient	Impressive	Personal	Unattractive
Comfortable	Effortless	Incomprehensible	Poor quality	Uncontrollable
Compatible	Empowering	Inconsistent	Powerful	Unconventional
Compelling	Energetic	Ineffective	Predictable	Understandable
Complex	Engaging	Innovative	Professional	Undesirable
Comprehensive	Entertaining	Inspiring	Relevant	Unpredictable
Confident	Enthusiastic	Integrated	Reliable	Unrefined
Confusing	Essential	Intimidating	Responsive	Usable
Connected	Exceptional	Intuitive	Rigid	Useful
Consistent	Exciting	Inviting	Satisfying	Valuable
Controllable	Expected	Irrelevant	Secure	
Convenient	Familiar	Low maintenance	Simplistic	

Another method to measure satisfaction would be post-test questionnaires containing questions or statements that participants respond using a rating scale or something similar. One standard method is questionnaires with Likert scales. Benedek and Miner (2002) claim that

participants tend to give similar responses, generally positive, to all questions when using questionnaires. Another problem is that participants' responses are restricted to the specific questions or statements defined by evaluation designers. An alternative method to measure satisfaction is to conduct interviews. While they can produce useful data, with some users it can be difficult to have a more negative feedback. Also they are time consuming to give and even more time consuming to analyze.

Tullis and Stetson (2004) conducted a comparative study to measure the effectiveness of five different feedback instruments in identifying user preferences for a financial website. The feedback instruments were adapted versions of:

1. System Usability Scale (Brooke, 1996) - a questionnaire developed at Digital Equipment Corp. that consists of ten statements and a rating on a five point scale of "Strongly Disagree" to "Strongly Agree".

2. Questionnaire for User Interface Satisfaction (Chin et al., 1988) - a questionnaire developed at the University of Maryland composed of 27 questions. Each question is a rating on a ten point scale with appropriate anchors at each end (e.g., "Overall Reaction to the Website: Terrible ... Wonderful").

3. Computer System Usability Questionnaire (Lewis, 1995) - a questionnaire developed at IBM composed of 19 statements and a rating on a seven point scale of "Strongly Disagree" to "Strongly Agree".

4. Words: Microsoft's Product Reaction Cards (Benedek and Miner, 2002) - a questionnaire based on the 118 words used by Microsoft on their Product Reaction Cards. Participants were free to choose as many or as few words as they wished.

5. Their own questionnaire (Tullis and Stetson, 2004) - a questionnaire that the authors have been using for several years in usability tests of websites. It is composed of nine statements (e.g., "This website is visually appealing") to which the user responds on a seven point scale from "Strongly Disagree" to "Strongly Agree".

One of the study conclusions is: "...Interestingly, on the surface at least, it appears that the Microsoft Words might provide the most diagnostic information, due to the potentially large number of descriptors involved." (Tullis and Stetson, 2004, p. 7).

Barnum and Palmer (2011) agree with this conclusion as they acknowledge that Product Reaction Cards have the ability to unlock information regarding the user's sense of satisfaction in a more revealing way than any other tool or technique they have tried. They have used the cards while studying systems with different levels of complexity. From their experience, the reason for the success of Product Reaction Cards is because it provides a way for users to tell their story, choosing the words that have meaning to them as triggers to express their feelings about their experience. In other words, participants have to build their own narrative.

This, on one hand, makes them more participative then answering pre-defined questions for which they have to assign a value. On the other hand, the broad set of descriptors involved (118) can lead to a more accurate feeling description because of its sheer size but also because participants can combine them in the way they feel more appropriate.

Three different methodologies of using Product Reaction Cards were found in literature. One of the methodologies consists in asking participants to select cards that reflect how they feel about their experience with the product (Benedek and Miner, 2002; Travis, 2008). Next, participants are asked to narrow their selection to their top five choices and then they are asked to explain their choices, picking their story-line and its outcome from the prompt provided by the word or phrase on each top five card.

The second methodology was developed to evaluate four design options in three stages (Williams et al., 2004). In stage 1 individual measures are gathered without discussion. Participants are instructed to review different design options spending one minute in front of each design in each format. They are asked to complete a form indicating their choice of the best design and also to rank the other choices. In stage 2, participants are grouped with others who had made the same choice in stage 1, and engage in several activities like marking up pages for each design. Following this task, they are asked to write down three words they think best describes each of the designs and then complete a Likert type questionnaire. The third stage is about achieving group consensus. Each group of stage 2 is asked to reach consensus on the top three cards that reflect their preferred design using Product Reaction Cards and to provide a brief explanation on the reason for each card chosen.

The third methodology found in literature was conceived to study desirability at eBay and Yahoo! by creating cards that matched brand values (Rohrer, 2008). In the desirability studies conducted at eBay, pairs of opposite words or phrases were used to assess responses. Participants were asked to select one element of each pair such as: uninteresting-interesting; makes me feel unsafe-makes me feel safe; forgettable-captivating; cluttered-clean. For Yahoo!, participants were instructed to select cards to express their responses to various designs. They were shown different visual designs for Yahoo! Personals, and asked to select the cards that best matched their responses to each design. Participants were then interviewed to understand their reasons for selecting the cards they did. This method was adapted for use in a quantitative study in which various images were embedded into a survey. Participants were asked to respond by choosing a word or phrase from a list that matched their feelings about each design.

6.1.1 Methodology

In this work, participants were required to test the proof-of-concept prototype and systems currently available for the public. The purpose of testing the systems currently available was to use them as a sentiment baseline. Search using the prototype was only via a natural language

interface, in order to provide similar search methods as the ones available in currently available systems (see Figures 6.1, 6.2, and 6.3). So, other search methods as pins on a map were not available. The interface layout of the prototype was a search box in a website mimicking the search mechanisms provided by municipalities.

The evaluation session started with an automated presentation explaining the session objective, the tasks to be performed and their sequence. After presentation, each participant received a random identifier to anonymize the session, and there was no registration of which participant received which identifier. The identifier was used to record participants' inputs anonymously and also to define the system order: half of identifiers started the experiment using systems currently available and the other half started the experiment using the prototype. The motivation behind the swap in system order was to remove bias towards the first or second search method tested.

Each participant spent, individually, 15 minutes searching information with each system, in the order defined by the respective identifier. The information to be searched was randomly selected from a pool of 35 questions. Participants were instructed to spend no more than 5 minutes with a given question and were allowed to skip questions at will.

At the end of the session, participants' inputs were collected using a methodology similar to the one described in Travis (2008). First, they could select as many cards as they wanted to describe each system. The cards were selected in a website containing 11 pages where each page showed 10 random cards at a time, without repeating cards. A 12th page showed the remaining 8 cards that were not in the previous 11 pages. After selecting cards freely, participants had to select their top 5 choices for each system from the set of previously selected cards. This last step was made in specific web-page. No group discussion was made whatsoever.

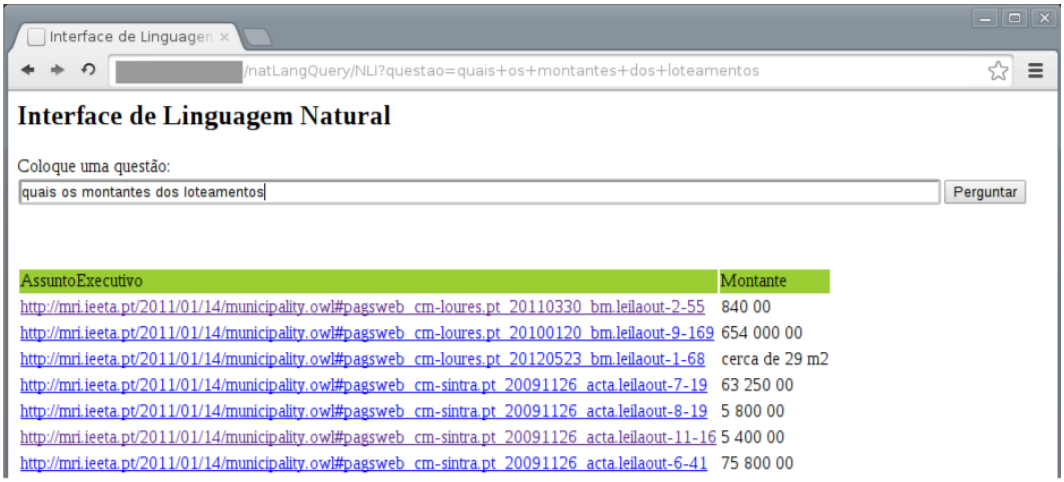


FIGURE 6.1: Screenshot of the prototype natural language interface with the question "quais os montantes dos loteamentos" which, informally, translates to "which are the allotments amounts". Visible answers are for the municipalities Loures and Sintra and, excluding the third, those numbers are amounts in euros. The third answer is an error that means "around 29 m^2".

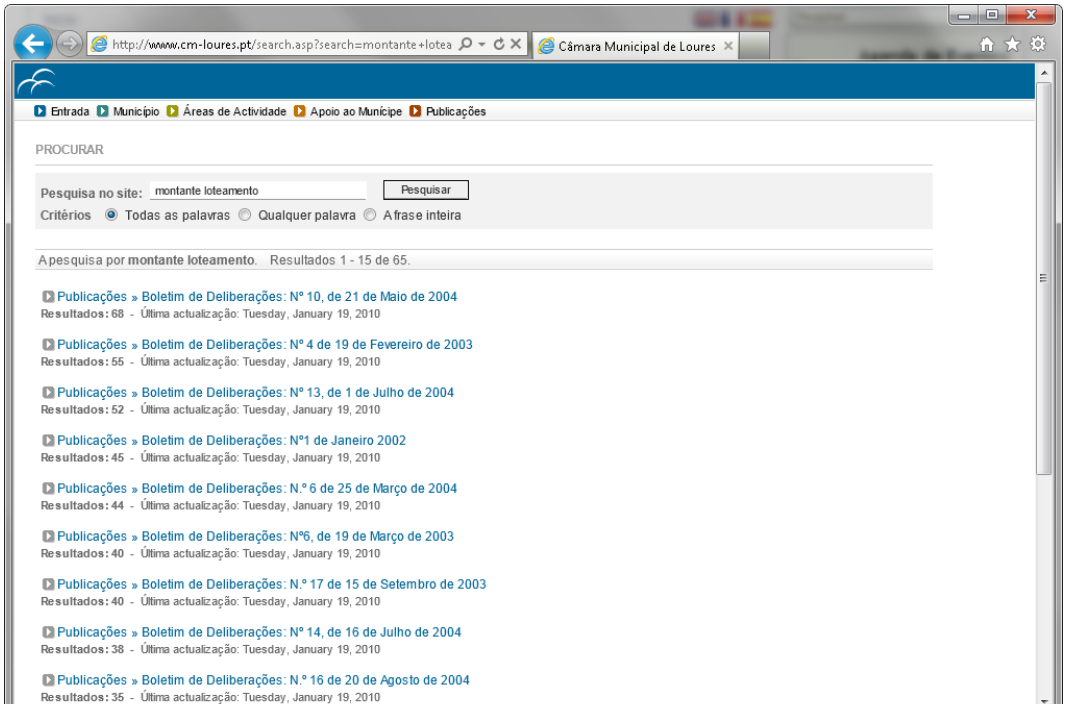


Figure 6.2: Screenshot of Loures municipality website with answers to the keywords "montante loteamento" which stand for "allotment amount". Answers are links to full documents and so it is necessary to open the documents to find the answers.

6.1.2 Experiment Setup and Participants

The data used in this test was obtained from public documents relative to the municipal mandate of 2009-2013 of the ten most populous Portuguese municipalities. The selected municipalities have a combined population of more than 2.5 million inhabitants, representing around 25% of Portuguese population (see Table 6.2). The type of documents retrieved were those referring the selected topics and included minutes of municipal assembly meetings, minutes of municipal executive meetings, and municipal bulletins. However, it was not possible to automatically acquire data from three of the selected municipalities and thus these municipalities were not included in this test. The excluded municipalities were: Porto, Vila Nova de Gaia, and Cascais.

In the case of Porto, it was not possible to obtain a list of the documents available. The way to access documents provided by the municipality website is a keyword search, or the latest document added to the system. With such mechanisms it is difficult to retrieve all documents. The difficulty with Vila Nova de Gaia and Cascais was of a different kind. In both municipalities all the documents were available in PDF and the problem was that the documents text was stored using an image format instead of text format. This introduces a significant amount of errors when trying to obtain the documents content, producing unreliable results and, in some cases, was not even possible to extract any meaningful content.

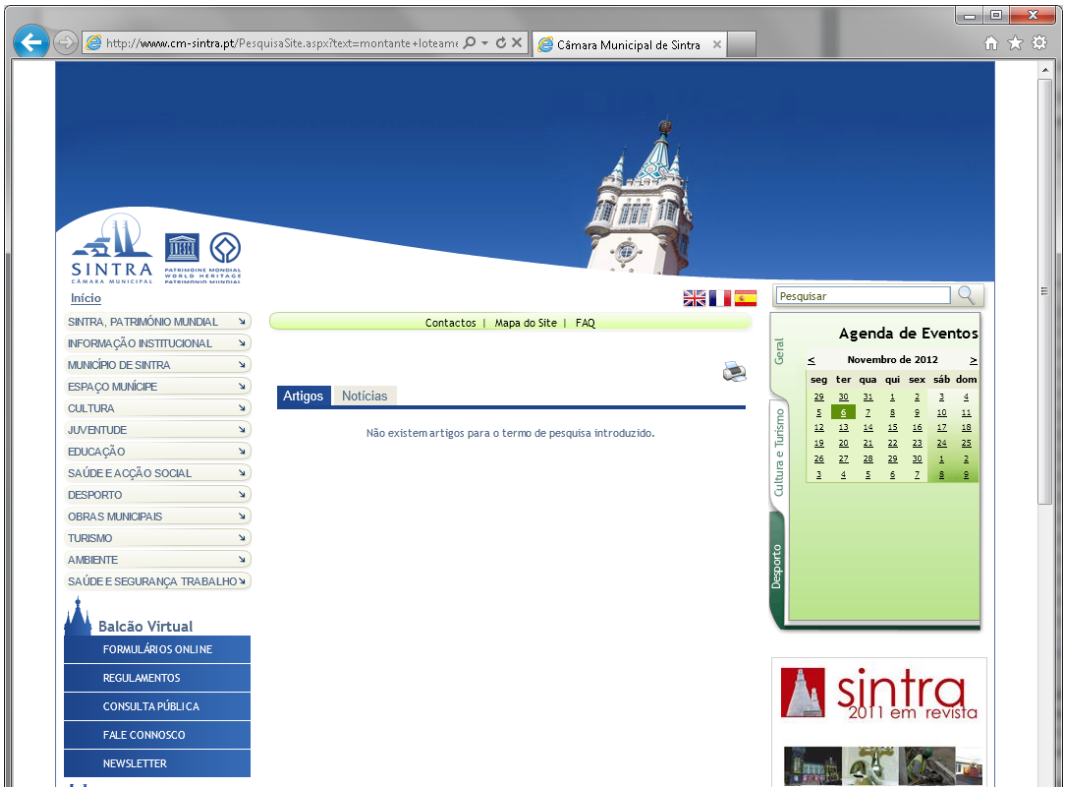


FIGURE 6.3: Screenshot of Sintra municipality website with no answers found to the keywords "montante loteamento" which stand for "allotment amount".

A random set containing 10% of the documents of each municipality was used to provide seed examples to train the semantic extraction models of the prototype. After training, the prototype automatically extracted information from the remaining 90% of documents. The extracted information was stored in a knowledge base and was accessible via natural language queries in a website.

Participants in this evaluation represented a target population familiar with Internet but not expert in ICT and not familiar with natural language technologies. A class of a European Level 4 qualification program on Office Administration and Translation was selected (in Portuguese, Curso de Especialização Tecnológica em Práticas Administrativas e Tradução). The class contained 22 students of which 14 were women. The first contact between participants and the authors of this work was during the evaluation session, and participants were not aware of this work before that session. All session arrangements were made between the authors and the program director.

6.1.3 Results

This section starts by analyzing all cards selected by participants with the objective of inferring the general attitude towards using each search alternative. The analysis also includes

Table 6.2: Top ten most populated Portuguese municipalities according to the preliminary results of Portuguese census 2011. Source INE (2012).

Municipality	Population	Docs used
Lisboa	547,631	354
Sintra	377,837	18
Vila Nova de Gaia	302,296	0
Porto	237,584	0
Cascais	206,429	0
Loures	205,054	79
Braga	181,474	27
Matosinhos	175,478	88
Amadora	175,135	101
Almada	174,030	112
Total	2,582,948	779

the cards that were not selected to describe each system. The section ends focusing on the collection of top 5 cards that participants were asked to choose to best describe their experience. Restricting participants' choices to 5 cards forces them to choose the more meaningful cards and eliminates possible bias caused by different levels of expressiveness among participants since each participant contributes with the same amount of cards for the result set.

The overall number of cards selected by participants show that they were more expressive with the proof-of-concept prototype (401 cards) than with systems that are currently available (281 cards). The only reason found to explain this difference is that the novelty factor introduced by the natural language interface compelled participants to provide a more detailed description of their feelings. Figure 6.4 shows how many participants selected a number of cards belonging to the intervals $[0; 10[$, $[10; 20[$, $[20; 30[$, and $[30, 40[$ (no participant selected more than 39 cards). It is visible that 8 participants selected less than 10 cards in the case of municipalities' systems, and 4 in the case of the proof-of-concept prototype. Also, no one chose 30 or more cards to describe municipal systems whereas for the prototype 5 participants selected between 30 and 39 cards. The most frequent amount of cards selected belonged to the category of 10 to 19 cards for both search alternatives. This can be due to the cards being dispersed in 12 website pages as the participants rarely left a page without any card selected or chose more than 2 cards in one page.

Regarding what participants felt when searching information, see All column in Table 6.3, 77.6% of all cards selected for the prototype had a positive meaning (☺), showing a positive attitude towards the prototype. For the municipalities' systems, the positive percentage was lower at 51.6%, which can be interpreted as a neutral attitude. This neutral attitude can be explained by participants being accustomed to use keyword based search systems, the type of search used by all municipalities tested.

Participants that used the prototype after experiencing the current search systems of municipalities (see column MUN first) selected a bigger proportion of positive cards to describe

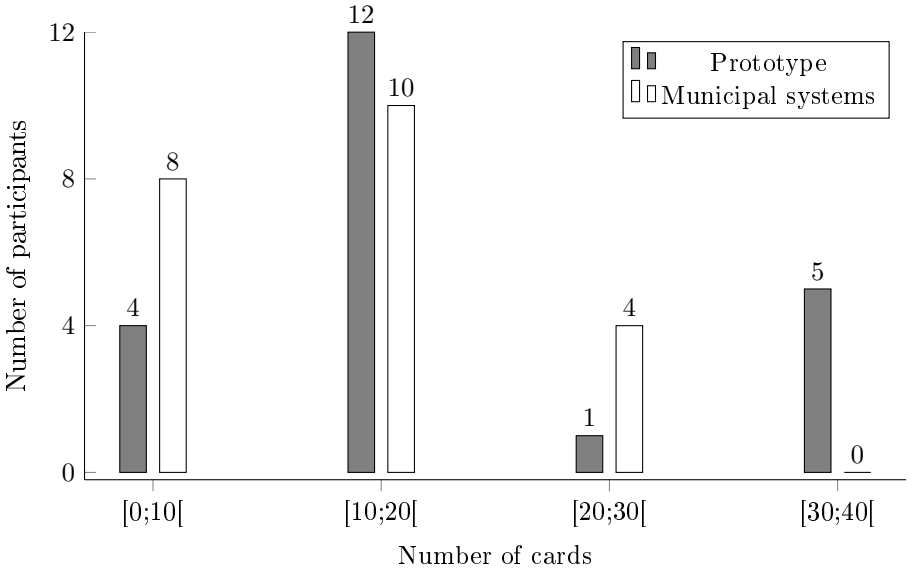


Figure 6.4: Number of cards selected by participants for both the prototype and the systems currently used by municipalities.

the prototype: a proportion of 82.4% cards were positive when the prototype was used in second place against 71.3% when participants started the experiment using the prototype (PROT first). As no interview was made after the experiment, it is not possible to know exactly the reason for this phenomenon. It can be a normal difference as participants did not have to agree with each other, or it can be due to the enthusiasm of experiencing a novel system, the prototype, just before selecting the cards. Nevertheless, the prototype always motivated a bigger proportion of positive cards than the systems currently available.

Besides observing the proportion of positive/negative cards selected altogether was also sought how many participants selected more positive than negative cards and vice-versa. The number of participants that had a positive attitude, meaning that selected more positive cards than negative, is also presented in Table 6.3, Participants section. Not all participants were accountable in this measurement because: (^{1}) one participant selected the same amount of positive and negative cards for the prototype; (^{2}) two participants did not select any cards to describe the currently available systems.

Overall it is visible that around 76% of participants selected more positive than negative cards for the proof-of-concept prototype and that 60% of participants did the same for the currently available systems (see All column, Participants section). Also, more participants had a positive attitude describing the prototype when they used it after experiencing the currently available alternatives (see MUN first). This is the same phenomenon observed for the total amount of cards selected. Table 6.3 also shows that participants with a positive attitude were the same or more than participants with a negative attitude, and that the prototype had at least the same amount of participants with positive attitude than the systems currently

available.

Table 6.3: Sentiment distribution of all cards selected. Sentiments were categorized as positive (☺) or negative (☹) according to the words and phrases of the cards. Results are divided by participants that started the experience using the prototype (PROT first) and the ones that started using municipalities' search systems (MUN first). Participants excluded: ([1]) one participant selected the same amount of positive and negative cards; ([2]) two participants did not selected any cards.

	All		**PROT first**		**MUN first**	
	PROT	MUN	PROT	MUN	PROT	MUN
Cards						
☺	311 (77.6%)	145 (51.6%)	124 (71.3%)	73 (55.7%)	187 (82.4%)	72 (48.0%)
☹	90 (22.4%)	136 (48.4%)	50 (28.7%)	58 (44.3%)	40 (17.6%)	78 (52.0%)
Total	401	281	174	131	227	150
Participants						
☺	16 (76.2%)	12 (60.0%)	7 (70.0%)	7 (70.0%)	9 (81.8%)	5 (50.0%)
☹	5 (23.8%)	8 (40.0%)	3 (30.0%)	3 (30.0%)	2 (18.2%)	5 (50.0%)
Total	21[1]	20[2]	10[1]	10[2]	11	10[2]

The last indicator taking into account the generality of cards selected by participants was the set of cards not selected by any participant. Results are presented in Table 6.4. Among others, for both search alternatives no one selected Cutting-edge, Disruptive or Uncontrollable which possibly means that everything behaved as all people expected. Also, no one described the systems as Empowering. Considering that cards like Useful and Effective were selected, this could mean that accessing information is not enough to make people feel empowered or that the information available was not empowering. A hypothesis to be researched is if the same holds true when integrating information provision systems with systems that allow to act upon that information.

Cards that were not selected to describe the proof-of-concept prototype, adding to the ones not selected for both alternatives: Integrated - it is natural as the prototype is not as well integrated as a final product; Irrelevant - it is good news that no one felt that the prototype is irrelevant, meaning that it can have a relevant role; Rigid and Familiar- participants found the approach different from what they already know and not rigid.

Some cards not selected for the municipalities' system were: Attractive and Creative - which is natural as people in general are used to such systems; Efficient, Time-saving and Valuable - which can be due to the large amount of results returned for each search and, adding to it, retrieved results are usually full documents and not relevant snippets of information.

Top 5 Choices

In a first stage, participants selected all cards that felt appropriate. In a second stage, participants were requested to restrict their choices to a set of 5 cards. Due to one participant just having selected 4 cards for the prototype in the first stage, the total number of cards in the

Table 6.4: List of cards not selected by participants for both alternatives, just for the proof-of-concept prototype (PROT) and just for municipalities current search systems (MUN).

Alternatives	Cards Not Selected by Participants
Both	Busy, Cutting-edge, Disruptive, Empowering, Entertaining, Patronizing, Personal, Uncontrollable.
PROT	Distracting, Exceptional, Familiar, Integrated, Irrelevant, Overbearing, Rigid.
MUN	Attractive, Clean, Connected, Creative, Efficient, Exciting, Expected, Impressive, Inspiring, Intimidating, Inviting, Novel, Responsive, Time-saving, Valuable.

second stage for the prototype was 109 instead of 110. For the municipalities search systems, in the first stage, two participants did not select any card and another one just selected 4 cards. As such, in the second stage the total number of cards for the municipalities systems was 99 instead of 110.

The analysis to the top 5 choices starts by checking if the general attitude towards the search alternatives remains similar to the attitude revealed when considering all cards. Figure 6.5 depicts the proportion of positive cards selected for both search alternatives, considering the top 5 choices and all cards selected. It is visible that the proportion of positive cards of the top 5 choices is similar to the ones found in the full set of selected cards. The set of top 5 cards selected by participants contained around 73% of positive cards for the prototype and around 51% for the municipalities search systems. These results are also in Table 6.5 (see All column, Cards section) and detailed by the order that participants used the search alternatives (PROT first or MUN first). It is visible that when the prototype was used after the current search systems it obtained a bigger proportion of positive cards. For the current search systems the proportion remains roughly around 50% independently of the usage order.

Considering the number of participants that selected more positive than negative cards (section Participants in Table 6.5) it is now possible to count all participants as no tie was observed. Participants that selected more positive than negative cards in their top 5 choices were again 16 globally, being 8 regardless of the prototype being used in the first or in second place. Checking the answers of the participant not counted when all cards were considered, due to the tie between positive and negative cards, it was found that its top 5 selection was 3 positive and 2 negative cards.

Regarding the municipalities search systems, the number of participants that made their top 5 selection with more positive than negative cards was 10. This shows that when participants were asked to prioritize their choices, they were less positive about municipalities search systems, as 12 participants selected more positive than negative cards when all cards were considered. This decrease was just observed when municipalities' systems were used after the prototype (compare Participants section and PROT first column in Tables 6.3 and 6.3).

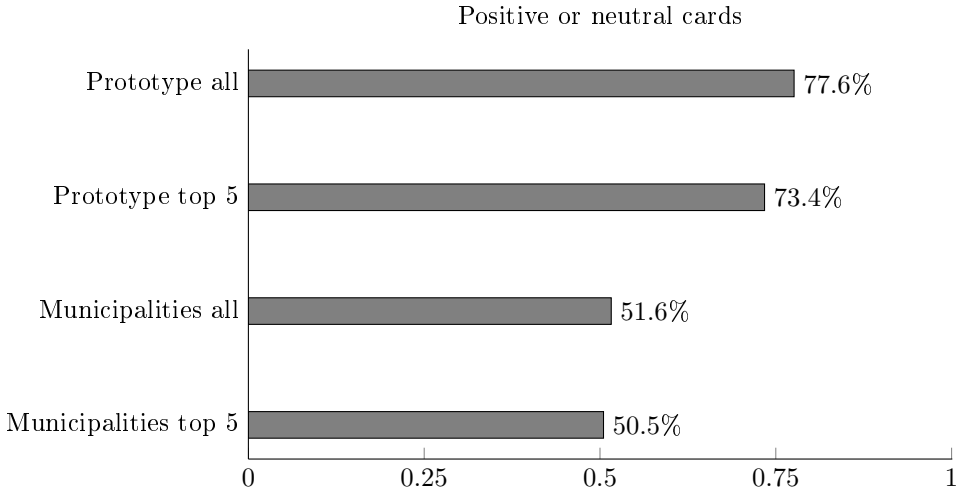


Figure 6.5: Proportion of positive and neutral cards selected for both search alternatives, considering the top 5 choices and all cards selected.

Looking to the Words and Phrases Selected for Top 5

To understand what participants felt using both search alternatives it is necessary to know how they describe them. For this was measured the frequency of cards that were part of the top 5 choices of three or more participants. Results can be seen in Figures 6.6 and 6.8.

For the prototype, the card most selected was Accessible (see Figure 6.6). Selected by nine participants, it means that the usage of natural language queries didn't bring major concerns and that participants felt comfortable using the prototype. However, with Confusing being the second most selected card, it also means that some participants might have struggled to perform the queries or to interpret its results. Further studies should be conducted to find if the main cause of confusion for some participants were the queries or the replies using information snippets, or both.

The cards Effective, Efficient, Inviting, and Useful were selected by four participants, becoming the third more frequent cards. All cards denote core values of information search systems. This selection is even more relevant as participants did not use these cards to describe the currently available alternatives. This does not mean that the current alternatives do not have those values, it means that participants felt the prototype as more effective and efficient, and thus useful, when comparing with current search systems of municipalities.

The prototype was also described positively as Advanced, Clear, Fast, and Straightforward and negatively as Dull and Unattractive. Both negative attributes can be due to the visual appeal not being a priority in this early stage of development. When further developing the attractiveness of the system, one must check if attributes as Clear and Straightforward are not lost.

All top 5 cards selected by participants are in Figure 6.7, including all cards selected by

Table 6.5: Sentiment distribution of participants' top 5 choices for the prototype and the systems currently available. Sentiments were categorized as positive (☺) or negative (☹) according to the words and phrases of the cards. Results are divided by participants that started the experience using the prototype (PROT first) and the ones that started using municipalities' search systems (MUN first). ([1]) participant just selected 4 cards; ([2]) participant did not selected any cards.

	All		**PROT first**		**MUN first**	
	PROT	MUN	PROT	MUN	PROT	MUN
Cards						
☺	80 (73.4%)	50 (50.5%)	39 (70.9%)	26 (52.0%)	41 (75.9%)	24 (49.0%)
☹	29 (26.6%)	49 (49.5%)	16 (29.1%)	24 (48.0%)	13 (24.1%)	25 (51.0%)
Total	109[1]	99[1 2]	55	50[2]	54[1]	49[1 2]
Participants						
☺	16 (72.7%)	10 (50.0%)	8 (72.7%)	5 (50.0%)	8 (72.7%)	5 (50.0%)
☹	6 (27.3%)	10 (50.0%)	3 (27.3%)	5 (50.0%)	3 (27.3%)	5 (50.0%)
Total	22	20[2]	11	10[2]	11	10[2]

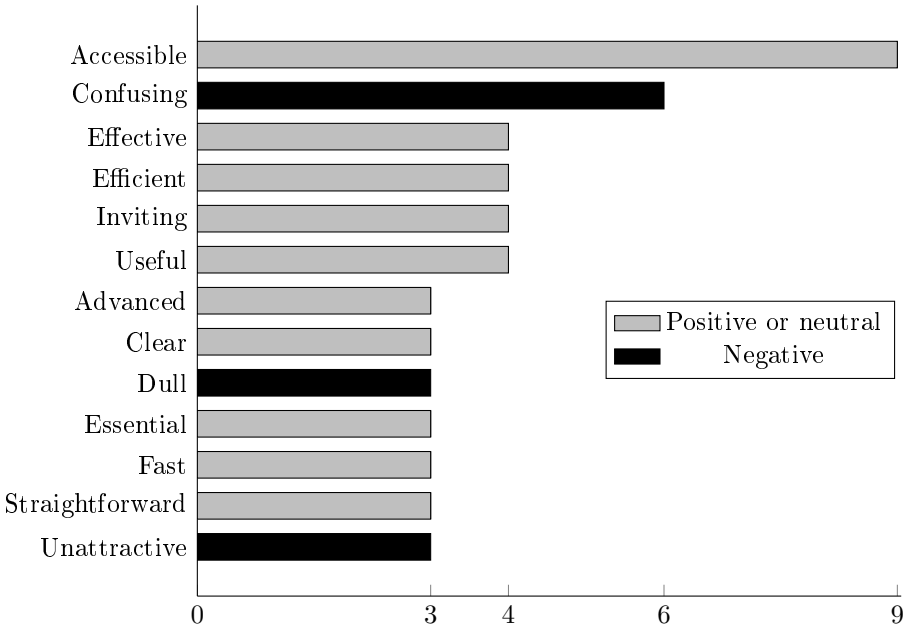


Figure 6.6: Frequency of cards selected as top 5 for the prototype, by three or more participants.

less than three participants and that were not included in Figure 6.6. In Figure 6.7, the font size of the text is directly proportional to that card frequency. As such, Accessible is the word with larger font, followed by Confusing and so on and so forth.

For the municipalities search systems, the card most selected was Time-consuming (see Figures 6.8 and 6.9). Selected by nine participants, it means that participants felt that it is too slow to find results. Slow, which is a similar feeling, was the second most selected card. The time took by the search systems in producing answers was not measured. However, participants informally said that everything went normally during the experiment, which can

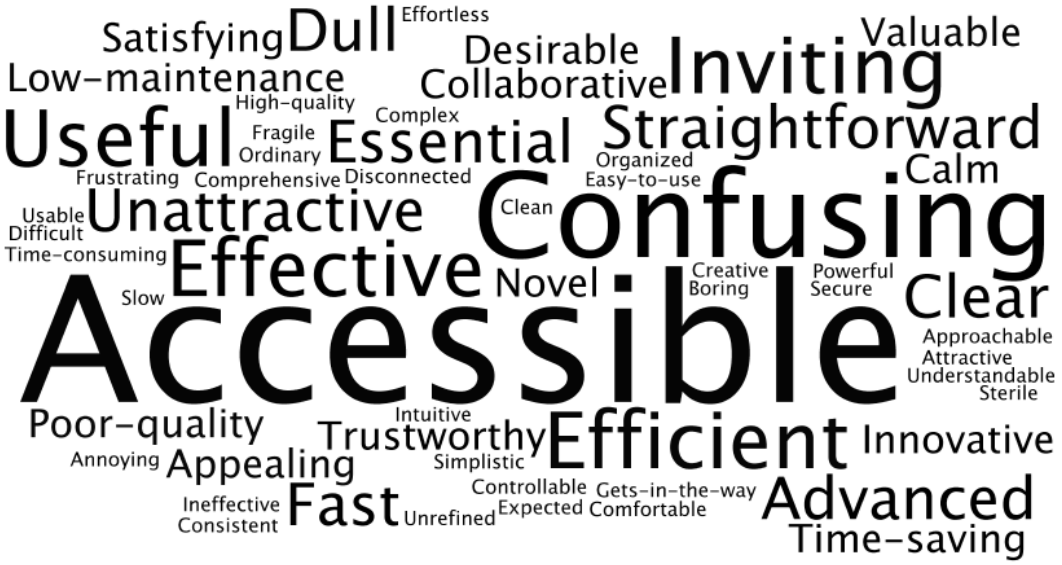


Figure 6.7: Cards selected as top 5 of the prototype. Larger font size means that the card was selected by more participants.

be assumed as no significant time lag was observed. The most probable cause is how results are provided. Most, if not all, municipalities search systems results are sets of documents that contain the keywords queried somewhere in its texts. This makes necessary to search the exact information in each document of the set. This is more time consuming than just have a snippet containing the queried information.

Also, Satisfying was selected by six participants as a key feature. This is only natural as participants should be familiar to using keyword search systems, and so they feel that these systems meet their expectations. The same applies to Accessible and Dull being the fourth and fifth most selected cards, respectively.

The cards Essential, Trustworthy, Confusing, and Stressful were selected by three participants. The first two cards mean that participants felt that they can rely on such systems, and again these cards are associated with familiarity on using these systems. The last two cards mean that the experience was enervating for some participants. As keyword based search is the most common method for information search on the web, these feeling might be originated by the way answers were provided. Having large sets of documents as result makes necessary to inspect each document until a satisfying answer is found. If an answer is not found, it is necessary to try another query and traverse the result set again.

6.1.4 Discussion

Satisfaction was elicited using Microsoft Product Reaction Cards (Benedek and Miner, 2002). To be able to derive meaning from the results it was necessary to establish a referential for comparison, otherwise would be difficult to know if participants' attitude towards the

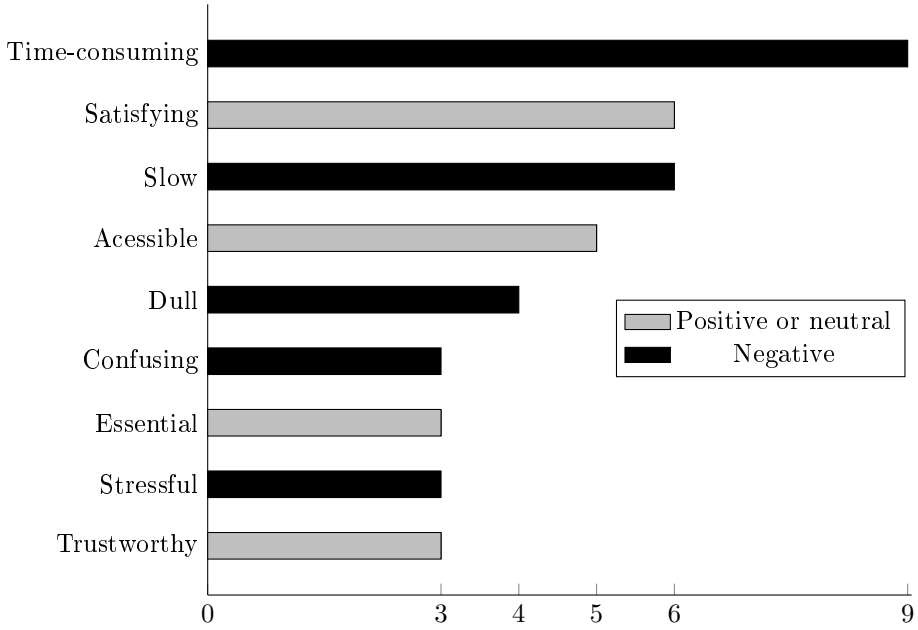


Figure 6.8: Frequency of cards selected as top 5 of search systems of municipalities, by three or more participants.

Figure 6.9: Cards selected as top 5 of search systems of municipalities. Larger font size means that the card was selected by more participants.

prototype was better than expected, worse than expected, or as expected. The referential used in the experiment were the currently available municipalities search systems. They are a good referential because:

1. They establish a baseline as they share the same technology and use a widespread search approach: keyword based search. As such, participants should be comfortable with these systems and should have a somehow neutral attitude.

2. They are relevant examples of services that the prototype aims to improve upon.

Results show that participants favored the prototype over the municipalities search systems. When they could chose cards at will, the percentage of positive cards was over 70% for the prototype and, at most, 55.7% for the municipalities search systems. Moreover, the number of participants that selected more positive than negative cards for the prototype was, in the worst case, as many as for the municipalities search systems.

When participants were forced to choose at most 5 cards to describe each search alternative, the percentage of positive cards remained over 70% for the prototype and was at most 52% for the municipalities search systems. Considering these top 5 cards, the number of participants that selected more positive than negative cards for the prototype was always higher than for the municipalities search systems.

Effective and efficient were two of the key features selected by at least four participants for the prototype and not for the other systems, even when participants were free to select as many cards as they wanted. Adding to these two properties, Useful was also only selected to describe the proof-of-concept prototype.

On a negative note, some participants also found the prototype to be Confusing, Dull and Unattractive. The first two negative cards were also used to describe the other systems involved in the tests and they should be taken into account in future developments. The most concerning card is Confusing as it possibly prevents people to effectively use this kind of systems. It can be associated to one of the key features of the prototype: the use of a NLI. It brings more expressiveness that is helpful to accurately specify the information to be found but it can be confusing if users do not understand or trust the system. A solution is the existence of an alternative and simpler user interface.

The usability assessment revealed that participants felt the proof-of-concept prototype as an improvement over what currently exists. Despite its early stage of development, the prototype showed that it is possible to improve government information provision by using NLP technologies.

6.2 Information Extraction Performance

The IE part of the developed prototype contains many components that have impact on the global performance. The performance of all components was evaluated individually before its inclusion in the system. All results obtained with this individual evaluation were in line with the ones reported in literature and will not be addressed here. Here the objective is to measure the performance of the system end to end, which means retrieving relevant information from natural language texts both in terms of the amount of information detected and the accuracy of the retrieved information.

One difficulty when comparing with state of the art systems is that most of the best performing systems reported work with English language and the prototype works with Portuguese. Moreover, for Portuguese, the general domain systems extract and classify named entities but do not extract relations between them. If we restrict the comparison to named entity extraction and classification, this implies that the component evaluated is essentially REMBRANDT, which performance was already measured by its author in Cardoso (2008). The experiment described here measures the prototype performance doing the complete IE

task and compares it with state of the art systems for English. This comparison is just an indicator of performance because it is not possible to perform direct comparisons due to the different working contexts.

Results are presented according to standard metrics frequently used in the evaluation of IE systems (Makhoul et al., 1999): precision, recall and F-measure. Precision is the ratio between the number of correct or relevant findings and the number of all findings of the system (see eq. 6.1), recall is the ratio between the number of correct or relevant findings and the number of expected findings which are the total amount of relevant facts that exist in the documents (see eq. 6.2). F-measure is the weighted harmonic mean of precision and recall (see eq. 6.3), commonly calculated as F_1 which is when β is equal to 1 (see eq. 6.4).

$$precision = \frac{number\ of\ correct\ findings}{number\ of\ findings} \tag{6.1}$$

$$recall = \frac{number\ of\ correct\ findings}{number\ of\ expected\ findings} \tag{6.2}$$

$$F_\beta = (1+\beta)\ \frac{precision \times recall}{\beta^2 \times precision + recall} \tag{6.3}$$

$$F_1 = 2\ \frac{precision \times recall}{precision + recall} \tag{6.4}$$

A difficulty when computing these measures is that it is necessary to know all the relevant findings of the documents, specifically when calculating recall, and thus F-measure. This implies having someone reading all documents and annotating the relevant parts of texts, which is a time consuming task. Ideally, the annotation should be performed by more than one person and followed by group consensus about which annotations are the correct ones. It is possible to find some sets of documents already annotated by linguists, named golden collections, but none was suitable for the scenario defined for testing the prototype: Portuguese municipalities information provision.

Performance measurements were made during the prototype development. As the IE component of the prototype did not have any change after the latest measurements, they were not repeated using the data gathered for the usability assessment. As such, experiments presented here were made before the usability assessment and are already published in Rodrigues et al. (2011b) and in Rodrigues et al. (2011a).

6.2.1 Methodology

Experiments were conducted to extract information about three topics of municipalities public documents: subsidies granted, building permits requested, and protocols with other institutions. A web crawler obtained all documents available in the Portuguese municipalities websites of the district of Aveiro. A total of 45.248 documents were retrieved, of which 2.277

are in PDF format. In general, the documents in PDF format are digital copies of official documents, and the remaining documents are pages about news, events agenda, contacts, staff information, etc.

Two random sets of 50 PDF documents each were selected: one for training the other for testing. The sets had no common documents and only PDF documents were selected because it is more likely to find relevant information in these than in any other type of documents. The training set was manually annotated using the annotator featured in the prototype. The annotations were about subsidies, build permits and protocols, and for each topic the following information was marked if available: requester, identifier, motivation, money amount, place, and deliberation (see Table 5.4 for details). The other set was used as the test set, having information automatically extracted by the prototype.

The test set contained 50 documents belonging to 7 municipalities. A person that read the 50 documents found information about 32 subsidies, 68 build permits and 41 protocols. This was considered the truth for this experiment. The output of the prototype for the test set was compared with the findings of the person. Information was considered found if the prototype detected one of subsidy, build permit, or protocol, even if it missed some facts like the requester or the money amount involved in the transaction.

6.2.2 Results

Results are presented in Table 6.6 and the number of subjects found for each municipality is the value between round brackets. The documents of one of the municipalities did not contain any relevant information for this test. The prototype was able to extract information about 14 of the 32 subsidies in the test set. This means that the recall value is 0.44 and, because all extractions were correct, the precision value is 1.00, leading to a F_1 equal to 0.61. Regarding build permits, from a total of 68 the system detected 67, meaning that the recall value is 0.99. Adding to the values presented in Table 6.6, for the municipality **a**, 4 build permits were incorrectly detected (see mark (^{1})). This implies that the precision was 0.94 and the resulting F_1 was 0.97. The system also detected 8 of the 41 existing protocols and also marked one protocol that did not exist (see mark (^{2})). Thus, for protocols, the recall value is 0.20, the precision is 0.89, and the F_1 is 0.32.

It is visible that the precision values are high and present less variability than the recall values. These high and stable values for precision mean that the system makes few mistakes when extracting information. This is important as the system should not provide incorrect information.

Relative to the recall values, the observed values are low. These low values are in part due to the existence of enumerations. As these did not exist in the training set, the system just considered the first occurrence for every enumeration of the type "*... protocolos com as seguintes instituições:...*" (in English, "protocols with the following institutions") followed by

a list of institutions. This caused the low recall values observed for protocols, and to some extend also for subsidies, as each listed institutions, excluding the first one, was considered a protocol or subsidy not detected.

Table 6.6: Information detected by the system. For each municipality document set are presented: the total number of correctly detected information and, between round brackets, the total information found in those documents. Adding to these values are ([1]) 4 building permits incorrectly extracted and ([2]) 1 protocol incorrectly extracted.

	municipality									
	a	b	c	d	e	f	g	precision	recall	F_1
subsidy	0(2)	3(3)	4(11)	1(1)	1(1)	3(14)	0(0)	1.00	0.44	0.61
build permit	3(4)[1]	13(13)	47(47)	0(0)	0(0)	4(4)	0(0)	0.94	0.99	0.97
protocol	3(4)	3(3)	0(3)	0(0)	7(24)	2(7)[2]	0(0)	0.89	0.20	0.32
total								0.95	0.63	0.76

6.2.3 Discussion

Considering state of the art systems that are domain independent, most of them work for the English language. In this case, the global performance of the prototype (precision 0.95; recall 0.63) is comparable with the performance of state of the art systems: DBpedia reported precision from 0.86 to 0.99 and recall from 0.41 to 0.7 (Bizer et al., 2009), Kylin reported precision from 0.74 to 0.97 and recall from 0.61 to 0.96 (Wu et al., 2008), and YAGO/NAGA reported precision from 0.91 to 0.99 and did not report recall (Suchanek et al., 2007).

Although it is not possible to derive sound conclusions from this comparison, it is visible that results obtained are acceptable considering the current state of the art in information extraction, namely when extracting entities and respective relations.

6.3 Application Examples

This section contains simple and practical examples of how the proposed system can be deployed to create real applications. This section does not pretend to be exhaustive as it is not possible or practical to outline every possible usage scenario. The objective is to discuss examples of application, thus providing a better illustration of the potential systems like our prototype and stimulating further developments. Four different types of examples were defined: internal organization, transparency, citizenship, and activity indicator.

6.3.1 Internal Organization Example

For internal organization, it is useful, for example, to know which protocols were already signed, which subsidies were granted, which public constructions contracts are being executed, and the status of all construction processes. As such, the contribution for internal organization

is illustrated by querying how many protocols were made between the municipality and other institutions. Protocols are a more generic subject as they can be relative to subsidies, competency delegation or other subjects. It is possible to ask in which document(s) the information was found, which can be useful for locating the protocol statement. The SPARQL query is presented and explained in Table 6.7.

Table 6.7: SPARQL query relative to protocols and respective explanation on the right hand side.

Query lines	Explanation
prefix terms: <http://purl.org/dc/terms/> prefix foaf: <http://xmlns.com/foaf/0.1/> prefix rdf: <http://www.w3.org/.../22-rdf-syntax-ns#> prefix municip: <http://.../municip#>	Namespaces and prefixes
select **?mun ?entity** (count(distinct **?top**) as **?prots**)	Get municipality (**?mun**), requester (**?entity**), and number of protocols between them
where {	
?org rdf:**type** foaf:**Organization ;** foaf:**name ?mun**.	**?org** is an organization and its name is stored in **?mun**
?doc terms:**publisher ?org ;** foaf:**topic ?top**.	**?org** published the document (so it must be a municipality) and the topics found are stored in **?top**
?top rdf:**type** municip:**Protocol**.	Topic (**?top**) must be a protocol
optional {**?top** municip:**requester ?req**}.	Store protocol requester in **?req**
optional {**?req** foaf:**name ?entity**}.	Requester's name stored in **?entity**
}	

Results are presented in Table 6.8. The first column of the table indicates which municipality is involved in the protocol. The second column contains the other entity involved in the protocol. If the system failed to identify the entity, the field is marked as "–". The third column shows the amount of protocols between the two entities.

6.3.2 Transparency Example

An example contributing for government transparency is to know how different municipalities grant subsidies. As subsidies are a way of influencing society, it is important to have an overview of how subsidies are granted. This resume can be done automatically as the system is capable of making a resume of the subsidies asked to municipalities and if they were granted or not.

External structured information sources were used to acquire the municipalities coordinates. This allows for the query result to be fed to a web-page responsible to render it in a

Table 6.8: Number of protocols between a municipality and other institutions. The cases where the entity field is "–" the system was unable to detect the name of the institution that made the protocol with the municipality.

Municipality	Entity	Protocols
Arouca	instituto_politecnico_de_viana_do_castelo	1
Arouca	camara	1
Arouca	assoc._empresarial_do_concelho_de_arouca	1
Arouca	–	1
Oliv. Azemeis	–	8
S. J. Madeira	camara_municipal	1
S. J. Madeira	–	1
Anadia	assoc._human._bombeiros_voluntarios_anadia	1
Anadia	ministerio_da_educacao	1
Anadia	–	2

map using those coordinates and Google Maps. This way the information relative to a location is associated with the respective map point.

Figure 6.10 is a screen shot of the web-page showing information relative to three subsidies granted by Câmara Municipal de São João da Madeira. The first two results were found in the document cm.sjm.pt_1678 and the third one in cm-sjm.pt_5101. The system could not find which entities applied for the subsidies (the empty field requester) but it found the amount requested (the field moneyAmount with values 700, 970 and 875 respectively), and the decision of the municipality to grand the subsidies (field deliberation with value *atribuir*). Other query that can be made is the total amount of the subsidies granted.

6.3.3 Citizen Example

For citizens' concerns it was chosen building permits. In Portugal, territorial administration is one key area of municipal responsibility. An example of a citizen benefiting from this type of systems is when he/she accesses information about his/her build permit: why is it taking so long? Is it already approved?

Let's consider a citizen called Maria. Maria is a frequent name in Portugal. A keyword based query to the document test set returned 165 results. A semantic query about the build permits applied by citizens named Maria returned 2 results (see Table 6.10). Moreover, besides restricting the results by specifying the type of the searched information, it is possible to have a resume with the document where the information was found, the identifier that the municipality assigned to the process and the outcome so far. The SPARQL correspondent to the query about the building permit that resumes the information is presented in Table 6.9.

In Table 6.10, the first column shows the full name of the person, the document where the information was found is presented in the second column. In the third and fourth columns are presented the municipality identifier of the building permit process and the outcome so far, respectively. The outcome *solicitar* (request) means that the municipality is requesting


```
cm-sjm.pt_1678
    subsidy requested to Camara Municipal de Sao Joao da Madeira
    requester:
    moneyAmout: 700
    deliberation: atribuir

cm-sjm.pt_1678
    subsidy requested to Camara Municipal de Sao Joao da Madeira
    requester:
    moneyAmout: 970
    deliberation: atribuir

cm-sjm.pt_5101
    subsidy requested to Camara Municipal de Sao Joao da Madeira
    requester:
    moneyAmout: 875
    deliberation: atribuir
```

Figure 6.10: Screenshot of a website page rendering a SPARQL query output in a map. The query is relative to the transparency example: How many subsidies each municipality grants? The information below the map is relative to the location indicated by the arrow.

for more documentation and this is the reason why the building permit process was not yet completed.

6.3.4 Other Concerns such as Activity Indicator

Besides specific topics about subjects related to municipalities, these type of systems can also be useful as tools for other purposes as identifying society dynamics in order to support the definition of public policies and priorities.

As an indicator of social dynamics we queried the system about the location of all activities (with a known location) in a given municipality. The information was then rendered in a map. Figure 6.11 shows a zoomed part of a municipality to allow the discrimination of nearby

Table 6.9: SPARQL query about which persons applied a building permit.

```
PREFIX municip: <http://.../municipality.owl#>
SELECT ?person ?document ?id ?outcome
WHERE {
  ?proc rdf:type municip:BuildPermit; municip:requester ?pret.
  ?page foaf:topic ?proc; terms:title ?document.
  ?pret foaf:name ?person.
  OPTIONAL {?proc terms:identifier ?id}.
  OPTIONAL {?proc municip:deliberation ?outcome}.
  FILTER(REGEX(?person,"maria")).
 }
```

Table 6.10: Status of the building permits requested by citizens which name includes *Maria*. The outcome *solicitar* (request) means the citizen needs to present missing documents.

Person	Document	Id	Outcome
maria_adel...	cm-arouca.pt_ACTA_12_2009	12/09	solicitar
maria_hele...	cm-arouca.pt_ACTA_22_2008	153/2008	solicitar

locations. The selected part of the result shows three occurrences in the city Arouca, two in the neighborhood Tropeço, and one in the neighborhood Burgo, all belonging to Arouca municipality.

Figure 6.11: Map rendering part of Arouca municipality for which information was found. Locations are relative to institutions who applied for subsidies, have protocols with the municipality, or the location for which building permits were requested.

6.4 Summary

This chapter presented the methodologies used to evaluate the proof-of-concept prototype, and provided some examples illustrating how the prototype can be used in real applications. Two different methodologies were used, one aiming to assess the usability of the system, and another aiming to measure the performance in extracting information from natural language texts in Portuguese.

The usability assessment involved having people, not related to this work, trying to obtain answers to some possible concerns using the prototype and using existing alternatives. Participants' impressions were collected using Product Reaction Cards, and results show that participants felt that the prototype was more effective, efficient and useful, among other characteristics, than existing alternatives.

The IE performance measurement implied a manual inspection to the IE module output of the prototype. Results show that, in general, the prototype is very precise having reached an average precision around 95% using public minutes of municipal council meetings. The recall value was lower at 63% which is comparable with what is achieved by domain independent state of the art systems for English. However, the results also indicate the need for several improvements, such as the effective detection of enumerations. Also, the prototype would benefit from the creation of more IE tools and resources in Portuguese.

The chapter also included some example applications that use the prototype capabilities at its core. These applications use the system mechanisms to read, process and utilize external resources to gather information and provide it with a semantic meaning. The example applications solve specific problems related to municipalities internal organization, transparency and citizens' concerns. By presenting these simple applications of the system we present examples of the possible real applications and how NLP technologies can be connected to develop more comprehensive solutions for the use of society.

7 Conclusions

This final chapter presents the milestones that have been pursuit and achieved over the course of this research, in Section 7.1. In Section 7.2 it is argued that the thesis statement was verified and are presented the main strengths and weaknesses of the conceptual model and respective prototype. Section 7.3 contains suggestions about possible directions for future research and development, and the chapter ends in Section 7.4 with some final considerations.

7.1 Work Overview

Some tasks such as literature review were done periodically over the course of the work. For clarity purposes, this overview is divided in three subsections and, in each one, just the more relevant activities will be mentioned. The subsections follow a chronological order and correspond to three different focus the work: (1) early study and exploratory tests; (2) conceptual model proposal and prototype development; (3) usability tests with external subjects.

7.1.1 Early Study and Exploratory Tests

The work started with literature review on how the inclusion of HLT in e-government systems could improve the information provided to society. The review had a broad scope and included articles from the HLT research area as well as from the e-government research area. Three important findings arose from this initial study:

1. There were not many initiatives that studied the inclusion of HLT in e-government

systems. This is a challenging problem which has been seldom addressed and requires more research. E-Government research initiatives usually use semantic technologies to specify, develop, and deploy services, and are rather centered in solving problems as interoperability and service integration.

2. The potential benefits of including HLT in e-government systems became clear: reduction of the digital divide; more intuitive human interfaces; communication comparable to traditional face-to-face dialogues; more information channels available; and more knowledgeable systems. These benefits are particularly relevant in e-government because government must serve all population.

3. How the system is designed is of crucial importance. At this time our view was not completely clear on this but it was visible that, besides political will, some e-government initiatives could not thrive due to difficulties in using them after deployment.

The first two findings of this study are discussed in a position paper which is the first publication relative to this work (Rodrigues et al., 2010b). The third finding was not included in that publication as we needed to mature our view. The study of why and how the deployment impacts the success of e-government initiatives was done in parallel with the next tasks.

The first milestone was to build an exploratory prototype to put our view into practice. The prototype development started with some experiments on producing semantic information based on the content of public official documents. The first feature developed was the ability to search information based on locations. It involved detecting the geographic locations in documents, displaying those locations on a map, and accessing the original content by navigating on the map. A set of rules and a geographical ontology was used to detect the exact location of entities that have some kind of address clues in their name. The acquired information was stored in a relational database and could be accessed through a web page displaying a world map with marks on top of the locations found in texts. The map interface can be seen in Section 6.3.2 (Figure 6.10) and this experiment was reported in Rodrigues et al. (2010a).

A second iteration of the exploratory prototype was done with three objectives: (1) starting to define the software architecture of a complete proof-of-concept prototype; (2) add the ability to discriminate more information than just locations; (3) associate a formal semantic description to all information stored. These objectives motivated two important changes in the approach, which perdure in the final version of the proof-of-concept prototype:

1. A processing pipeline performing syntactic analysis replaced the set of rules originally built to detect geographical entities. The pipeline is composed by a POS tagger, a named entity recognizer, and a syntactic parser, developed by other research groups and made available as open source software. The POS tagger and the syntactic parser were trained to process Portuguese using corpus of Linguateca. The selected named entity

recognizer was developed for Portuguese and so its usage was straightforward. This way, the syntactic processing pipeline takes advantage of state of the art software and can be updated as new algorithms are developed.

2. A knowledge base stored as a triplestore replaced the relational database. The knowledge base conforms to an ontology specified with OWL. Using an ontology has the benefits of having the semantic explicitly defined plus the ability to perform semantic reasoning. The independence from lower level data models make ontologies suitable for specifying interfaces to independent, knowledge based services, and enables semantic interoperability among disparate systems.

A set of rules transformed the syntactic structures built by the pipeline to ontological individuals and respective relations to be stored in the knowledge base. As a custom set of rules was used, a change in the ontology most certainly would imply a software reprogramming.

This iteration of the prototype was tested with the task of extracting information relative to municipal subsidies. Experiments focused on some particular subjects but the approach was versatile and generic. The extracted information included the amount involved, the institution that received the subsidy, and the location of the head-office of that institution. The information of the knowledge base could be accessed by a map interface and by a SPARQL endpoint. The prototype description and the results obtained are reported in Rodrigues et al. (2010c).

7.1.2 Conceptual Model Proposal and Prototype Development

The next milestone was about generalizing the module that assigns semantic meaning to parts of syntactic structures, in order to use any given ontology. The idea was to allow changes in the semantic definition without need of software reprogramming. This was motivated by the understanding of how the prototype should be deployed. The view that was not so clear in the beginning of this work was now perceived with clarity: e-government information provision systems should be viewed as open platforms that provide data in a granular form, in formats that are open, structured and machine-readable. They also should be flexible to cope with the variety of governmental subjects relevant to society.

Data was already in an open and granular format by having a knowledge base conforming to an ontology. However, for the system to be usable, it is also necessary allowing its administrators to define data semantics at will. It is important to allow the system to be usable in a wide variety of scenarios, either controlled by the data provider, by the data consumer, or by a third party (see details on scenarios in Section 3.4). Moreover, it is important that the system seamlessly adapts to the way government operates. Governments are complex organizations designed to have checks and balances and so, systems aspiring to be used in such context, should not require changes in the way operations are performed. This adaptation

is achieved by having the system acquiring information from documents that governments already produce, and with the semantic defined by the system administrator.

The way e-government information systems should operate is related to how system is designed and not only with how it is implemented. For this reason, those findings motivated the proposal of a conceptual model for government information provision (see Chapter 3 for details). A published book chapter discusses the current restrictions and guidelines to future directions in e-government information provision systems (Rodrigues et al., 2013). In this book chapter it is argued that a conceptual model architecture as the one proposed here answers to e-government and societal (enterprise systems) concerns.

The objective was now to build a prototype following the proposed conceptual model. The syntactic pipeline remained the same of the previous version with the inclusion of a better sentence boundary detector. Three main features were added in this version: (1) an algorithm to train semantic extraction models based on seed examples of syntactic/semantic correspondences; (2) a graphical annotator to help the creation of seed examples; (3) the use of ontological rules to expand the examples: examples of one class are used as counterexamples of disjoint classes; examples of one class are used as examples of its superclass. The first experiments with this version are reported in Rodrigues et al. (2011b).

The conceptual model specifies that information can be accessed by machines and by humans. Machines can access information via a SPARQL endpoint. A NLI was added to allow search for any information since search via a map is not enough to access all information because some information is not bound to geographical locations. The NLI allows interaction using Portuguese permitting a detailed description of the information desired. The interface also behaves as a keyword based search if full sentences are not provided. This flexibility is important since government systems are meant to be used by all population, independently of their level of proficiency in using ICT. A journal article, Rodrigues et al. (2011a), provides a comprehensive description of the prototype including the newly added NLI.

7.1.3 Usability Tests with External Subjects

Having proposed a conceptual model aiming to improve the way government provides its information, and developed a conforming prototype, the final step was to know if, in a realistic scenario, the system is an improvement over what currently exists. So, the last milestone pursuit was to assess how citizens feel about using the developed proof-of-concept prototype. Several arrangements were necessary for this assessment:

- Search for a usability assessment tool that allowed a high degree of expressiveness for participants. There was some reluctance in using questionnaires because they restrict participants to the topics asked, and the authors of this work are not proficient in designing questionnaires. The tool selected was Product Reaction Cards (see Section 6.1 for details).

- Making the NLI interaction similar to the query tools available in currently existing systems. This implied some software re-engineering in order to integrate the NLI in a website page.
- Select average Internet users as participants. Selected participants were capable of using technology without being ICT experts, and were not aware of this project before the evaluation session.
- Build an online platform to anonymously collect participants' answers and translate the Product Reaction Cards to Portuguese.

7.2 Discussion

One of the outcomes of the work reported here is a working prototype. Its usability was assessed against some municipalities information search systems currently in use. Participants in the usability experiment did not have any particular skill that would benefit their experience using the prototype. They were not students of an ICT related course and were not aware of this work before the session where they have tested the prototype.

Results show that participants favored the prototype over the municipalities systems, finding the prototype an improvement over what currently exists. These results were achieved without requiring changes in the way government services operate. The documents used to build the knowledge base are public, and public services are required to produce and publish them.

The setup of the experiment and the results achieved allows us to conclude that the thesis statement is verified:

- Accessing information using the prototype is an improvement over what currently exists, according to participants. This is due to the usage of semantic and NLP technologies that provide the ability to understand texts and assign a semantic meaning to relevant parts of them. The possession of semantic information allows the prototype to better match users' requests as demonstrated by the information snippets.
- The prototype does not compromise government important concerns like the digital divide and information asymmetry. In the worst case it is as good or as bad as the currently available systems since the NLI can fallback for a keyword base search. In a normal usage, the NLI allows more expressiveness and possibly can allow engaging dialogues in future developments. The prototype also provides a SPARQL endpoint and thus third parties can develop other types of interfaces that can reduce digital divide (e.g. speech), taking advantage of systems like the implemented prototype.
- The prototype does not require a significant change in the way government services operate and produce information. All tests were made with public documents that

anyone can obtain from official websites. This is not saying that the prototype could not benefit from some changes: for instance, possibly the prototype performance (precision and recall) would improve if documents had already semantic annotations on them. This is saying that it is possible to improve information delivery even if the usage context remains the same.

However, the results obtained do not confirm one part of the thesis statement: this kind of systems increase society empowerment. Participants described the prototype as accessible, effective, efficient, and useful, among others, but not as empowering. There is no conclusive reason why participants did not select the card empowering to describe any of the systems tested because no questions were made about it. It is our belief that this happened due to one, or a mix, of the reasons:

- The type of information available and the questions asked did not made participants feel empowered;
- The Portuguese word for empowering, *empossante*, is not common and participants shied away from it.
- For people to feel empowered it is necessary to provide information together with mechanisms to act upon that information. If this is the reason, then information provision systems should be coupled with other systems. The architecture proposed and implemented in this work can be easily integrated with other systems as the SPARQL endpoint provides a standard access to the knowledge base.

Nevertheless, being the prototype described as accessible, effective, efficient, and useful, one can assume that the foundations for becoming empowering are already there.

7.2.1 Strengths of the Approach

In our view, the five most relevant strengths of the conceptual model, and of the prototype that implements it, are:

- The conceptual model is simple and clear, outlining what is essential in e-government information provision systems. Systems complying with this model can be implemented using other software tools than the ones used to build the proof-of-concept prototype.
- The modular architecture brings versatility as it can be used with different semantic representations and with different natural languages. Relative to the semantic representation, the only restriction is using an ontology defined in OWL. As for the natural language, it is necessary to have a syntactic representation using dependency structures.

- The use of NLP and semantic technologies allows manipulating contents to better match users' requests. By interpreting both the documents content and the users' requests it is possible to present information snippets that present a resume of the information requested.

- The use of standards to define data semantics and data communication enables semantic interoperability. This makes the model to be conforming the view of e-government 2.0, where government systems are platforms that provide data and services to be used by others as they see fit.

- Different data sources and different interfaces can be used. Data sources can be natural language documents as well as structured data sources. Interfaces can be maps, NLI or any interface that communicates via a SPARQL endpoint.

7.2.2 Weaknesses of the Approach

Some of the choices made for the conceptual model have negative effects that were not fully mitigated. Four weaknesses are:

- Allowing services - system administrators - to define the data semantics it is a benefit that haves its own risks. Some technical skills are required and, alongside the semantic definition, this feature implies the creation of seed examples. Two types of errors can be made in the setup process: (1) errors in the semantic definition; (2) errors in the definition of seed examples, whether in terms of incorrect annotations or in the amount of examples provided. After the system setup it is advisable to check if the system is performing as expected over an extended period of time.

- It can be difficult to know if the amount of seed examples is sufficient or if more examples are necessary. The amount of seed depends on the variability of sentences structures, which means that relevant information that has a greater variability of sentence types should need more examples. Also, annotating examples can be a time consuming and fastidious process.

- The performance of triplestores is not at par with the performance of relational databases used in currently available systems. Triplestore performance depends on the level of expressiveness of the ontology (slower with higher expressiveness) and on the amount of data in the knowledge base (slower with more data). This problem should decrease over time as triplestore technology evolves.

- Processing documents to find relevant information is a time consuming process. Syntactic parsing, in this case dependency parsing, is a time consuming task and, according to the approach followed, all sentences are parsed before it is decided if they contain

relevant information. As algorithms improve over time and computers become faster, this problem should decrease.

7.3 Future Directions

Several opportunities exist to improve and to extend this work. Relative to improvements, the future should be about maturing the technology. The proof-of-concept prototype demonstrated that the approach has potential. However, the prototype is not yet ready to be used as a production system and to replace current alternatives. Some key aspects should be taken into consideration:

- Build a set of ready to use configurations, including ontology and the respective semantic extraction models for some natural languages. This implies studying which are the relevant topics for a majority of scenarios.
- Further research about the statistical classifiers that decide the correspondence between syntactic structures and semantic content. These classifiers are a critical point with influence on how well the system is capable of acquiring information and thus they should deserve particular attention.
- Measure the limits regarding the number of ontological classes and relations that this approach is able to learn and use. The conceptual model does not impose a limit but it is only reasonable to believe that limits exist somewhere. These limits should be known.
- Improve component integration. As a research prototype, component integration prioritized logging and traceability of intermediate results and component status, in disfavor of speed and streamlined setups.

Relative to extending this work, some interesting possibilities are:

- Test the approach in other scenarios like, for instance, central government information. One interesting target information would be laws.
- Add user authentication mechanisms. As the proposed model was designed for handling public documents and public information, this feature was not required. The lack of user authentication prevents the inclusion of private information in systems strictly conforming to the model. For now, with the current design, it is not possible for a business to access private information related to its interaction with government. Examples of such information are proposals submitted for business contracts.
- Add a spoken language interface. People would be allowed to use their telephone and have the same level of service they would have using the web. Speech is the most natural

and easy existent interface, not only for people with special needs, but for people in general. Another advantage is that telephones are more popular than computers and more people feel comfortable with them. Adding speech capabilities would help to mitigate digital divide.

- Explore the addition of dialogue capabilities. By having a dialog, written or spoken, individuals that do not know which information to look for, would be able to explain it to discover a solution, instead of having to search until they find what they need. This feature would make the system more reactive and capable of taking the initiative in the conversation to clarify, ask for missing information, or make suggestions.

- Study the possibility of developing a document editor, or a plugin to some document editor, that would mark ontological classes and relations as texts are being written. Such tool would allow the document producer to correct the system when it is more focused in the document: at production time. This would imply a change in the conceptual model as it does not support content creation.

7.4 Epilogue

In this work were discussed the current challenges in e-government information, and was proposed a conceptual model able to detect and organize relevant information existing in unstructured natural language documents. Key features of this model are the ability to process natural language texts, to accept a knowledge domain defined by an ontology, to learn from examples how to extract information, and to provide multiple ways of accessing acquired information. The proposed model supports the extension for other natural languages by changing the NLP module, and without changing other modules. The model also features a structured information entry point to allow the inclusion of information sources that complement the information included in the natural language documents.

A proof-of-concept prototype was developed for Portuguese language. The software tools used and respective setup were described, and it was explained how the tools work together in order to have a complete and coherent system. Usability and performance tests made with publicly available documents from Portuguese municipalities show that the system is able to acquire useful, meaningful information. They also show the advantages of assigning a structure to the information by making it possible to display the information spatially, or to make semantic queries that return a reduced set of useful results, or even to make information summaries. These capabilities are possible when dealing with structured data, hence the importance of using this type of systems to assign structure to existing, official unstructured textual information.

Making relevant information more accessible is an important task for e-government. Governments produce large volumes of data and it is necessary to provide relevant information

to clients - citizens, businesses, other branches of government. An advantage of the proposed model is that it does not imply a significant change in the way government produces its documentation: the model fits the way government works and not the other way around. This is important because government services should not be diverted to other tasks outside their scope. We expect the proposed model to contribute by helping the design and development of computer systems that make available relevant information contained in natural language documents.

Bibliography

Adrian, B., Hees, J., Elst, L., Dengel, A., 2009. idocument: Using ontologies for extracting and annotating information from unstructured text. In: KI 2009: Advances in Artificial Intelligence. Vol. 5803 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 249–256.

Afonso, S., Bick, E., Haber, R., Santos, D., 2002. "floresta sintá(c)tica": a treebank for portuguese. In: Proc. of the Third Intern. Conf. on Language Resources and Evaluation (LREC). pp. 1698–1703.

Aires, R. V. X., Outubro 2000. Implementação, adaptação, combinação e avaliação de etiquetadores para o português do brasil. Dissertação de mestrado, Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo - Campus São Carlos.

Akerkar, R. A., Sajja, P. S., 2010. Knowledge-Based Systems. Jones and Bartlett Publishers, Sudbury, MA, USA.

Allen, J. F., 2000. Natural language processing. In: Ralston, A., Reilly, E. D., Hemmendinger, D. (Eds.), Encyclopedia of Computer Science, 4th edition. John Wiley and Sons Ltd., Chichester, UK, pp. 1218–1222.

Aluísio, S. M., Pinheiro, G. M., Finger, M., das Graças V. Nunes, M., Tagnin, S. E. O., March 2003. The lacio-web project: overview and issues in brazilian portuguese corpora creation. In: Archer, D., Rayson, P., Wilson, A., McEnery, T. (Eds.), Proceedings of Corpus Linguistics. pp. 14–21.

Alves, C. D. C., Finger, M., September 1999. Etiquetagem do português clássico baseada em corpus. In: IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada. PROPOR99.

Andersen, K. N., Henriksen, H. Z., Medaglia, R., Danziger, J. N., Sannarnes, M. K., Enemærke, M., July 2010. Fads and facts of e-government: A review of impacts of e-government (2003–2009). International Journal of Public Administration 33 (11), 564–579.

Antoniou, G., van Harmelen, F., 2009. Web ontology language: Owl. In: Staab, S., Studer, R. (Eds.), Handbook on Ontologies, 2nd edition. International Handbooks on Information Systems. Springer-Verlag Berlin Heidelberg, pp. 91–110.

Apache Jena, 2013. Apache jena – a java framework for building semantic web applications. Accessed January 2013.
URL http://jena.apache.org/index.html

Ayres, Q. W., Kettinger, W. J., 1983. Information technology and models of governmental productivity. Public Administration Review 43 (6), 561–566.

Barnum, C. M., Palmer, L. A., 2011. Tapping into desirability in user experience. In: Alberts, M. J., Still, B. (Eds.), Usability of Complex Information Systems - Evaluation of User Interaction. CRC Press, Taylor & Francis Group, pp. 253–279.

Bechhofer, S., Goble, C., 2001. Towards annotation using daml+oil. In: K-CAP 2001 Workshop on Knowledge Markup and Semantic Annotation, Victoria BC.

Belanger, F., Carter, L., 2012. Digitizing government interactions with constituents: An historical review of e-government research in information systems. Journal of the Association for Information Systems 13 (5), 1.

Benedek, J., Miner, T., 2002. Measuring desirability: New methods for evaluating desirability in a usability lab setting. Proceedings of Usability Professionals Association, 8–12.

Berners-Lee, T., Hendler, J., Lassila, O., May 2001. The semantic web. Scientific American 284 (5), 34–43.

Bernstein, A., Kaufmann, E., 2006. Gino–a guided input natural language ontology editor. In: The Semantic Web-ISWC 2006. Springer, pp. 144–157.

Bick, E., 2000. The Parsing System "Palavras" - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Århus, Århus, Denmark.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S., September 2009. Dbpedia - a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web - The Web of Data 7 (3), 154–165.

Bohus, D., 2013. a list of spoken language interfaces. in Internet, accessed April 2013.
URL http://www.cs.cmu.edu/~dbohus/SDS/

Bohus, D., Rudnicky, A. I., 2009. The ravenclaw dialog management framework: Architecture and systems. Computer Speech & Language 23 (3), 332–361.

Bontcheva, K., Davis, B., Funk, A., Li, Y., Wang, T., 2009. Human language technologies. In: Davies, J., Grobelnik, M., Mladenic, D. (Eds.), Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery and Human Language Technology. Springer-Verlag Berlin / Heidelberg, pp. 37–49.

Branco, A., Silva, J., May 2004. Evaluating solutions for the rapid development of state-of-the-art pos taggers for portuguese. In: Lino, M. T., Xavier, M. F., Ferreira, F., Costa, R., Silva, R. (Eds.), Proceedings of the 4th International Conference on Language Resources and Evaluation. LREC2004. European Language Resources Association, pp. 507–510.

Brants, T., 2000. TnT - a statistical part-of-speech tagger. In: Proceedings of the 6th Applied Natural Language Processing. ANLC '00. Association for Computational Linguistics, Stroudsburg, pp. 224–231.

Brickley, D., Miller, L., 2010. Foaf vocabulary specification 0.98. Namespace Document 9.

Brill, E., 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics 21 (4), 543–565.

Brooke, J., 1996. Sus-a quick and dirty usability scale. Usability evaluation in industry 189, 194.

Buchholz, S., Marsi, E., 2006. Conll-x shared task on multilingual dependency parsing. In: Proceedings of the Tenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, pp. 149–164.

Buitelaar, P., Cimiano, P., Racioppa, S., Siegel, M., 2006. Ontology-based information extraction with soba. In: Proceedings of the International Conference on Language Resources and Evaluation. pp. 2321–2324.

Buitelaar, P., Ramaka, S., 2005. Unsupervised ontology-based semantic tagging for knowledge markup.

Cardoso, N., December 2008. REMBRANDT - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. In: Mota, C., Santos, D. (Eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, pp. 195–211.

Carlson, A., Betteridge, J., Hruschka, E. R., Mitchell, T. M., June 2009. Coupling semi-supervised learning of categories and relations. In: SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–9.

Carroll, A. B., Buchholtz, A. K., 2011. Business and Society: Ethics, Sustainability, and Stakeholder Management. South-Western Pub.

Chakravarthy, A., Ciravegna, F., Lanfranchi, V., November 2006. Cross-media document annotation and enrichment. In: SAAW2006 - Proceedings of the 1st Semantic Authoring and Annotation Workshop.

Chang, A.-M., Kannan, P. K., 2008. Leveraging web 2.0 in government. Tech. rep., IBM Center for The Business of Government.

Chatzidimitriou, M., Koumpis, A., February 2008. Marketing one-stop e-government solutions: the european onestopgov project. IAENG International Journal of Computer Science 35 (1), 74–79.

Chaves, M. S., Silva, M. J., Martins, B., October 2005. A geographic knowledge base for semantic web applications. In: Heuser, C. A. (Ed.), SBBD - 20^o Simpósio Brasileiro de Banco de Dados. UFU, Uberlândia, Brazil, pp. 40–54.

Chen, Y.-C., 2012. A framework for government 2.0 development and implementation: The case of u.s. federal government. In: Chen, Y.-C., Chu, P.-Y. (Eds.), Electronic Governance and Cross-Boundary Collaboration: Innovations and Advancing Tools. IGI Global.

Chin, J. P., Diehl, V. A., Norman, K. L., 1988. Development of an instrument measuring user satisfaction of the human-computer interface. In: Proceedings of the SIGCHI conference on Human factors in computing systems. CHI '88. ACM, New York, NY, USA, pp. 213–218.

Cimiano, P., Haase, P., Heizmann, J., 2007. Porting natural language interfaces between domains: an experimental user study with the orakel system. In: Proceedings of the 12th international conference on Intelligent user interfaces. ACM, pp. 180–189.

Cimiano, P., Handschuh, S., Staab, S., 2004. Towards the self-annotating web. In: Proceedings of the 13th international conference on World Wide Web. ACM, pp. 462–471.

Cimiano, P., Ladwig, G., Staab, S., 2005. Gimme'the context: context-driven automatic semantic annotation with c-pankow. In: Proceedings of the 14th international conference on World Wide Web. ACM, pp. 332–341.

Ciravegna, F., Chapman, S., Dingli, A., Wilks, Y., 2004. Learning to harvest information for the semantic web. In: The Semantic Web: Research and Applications. Lecture Notes in Computer Science. Springer, pp. 312–326.

Ciravegna, F., Dingli, A., Petrelli, D., Wilks, Y., 2002. User-system cooperation in document annotation based on information extraction. In: Gómez-Pérez, A., Benjamins, R. (Eds.), Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge

Management: Ontologies and the Semantic Web. Vol. 2473 of Lecture Notes in Artificial Intelligence. Springer Verlag, pp. 122–137.

Clarkson, G., Jacobsen, T., Batcheller, A., 2007. Information asymmetry and information sharing. Government information quarterly 24 (4), 827–839.

Codagnone, C., Wimmer, M. A., 2007. egovernment as a multidisciplinary research field. In: Codagnone, C., Wimmer, M. A. (Eds.), Roadmapping eGovernment Research: Visions and Measures towards Innovative Governments in 2020. pp. 12–14.

Colclough, G., Tinholt, D., November 2009. Smarter, faster, better egovernment: 8th benchmark measurement. Tech. rep., Prepared by: Capgemini, RAND Europe, IDC, SOGETI and DTi. For: European Commission Directorate General for Information Society and Media.

Collins, T., 2007. Only a third of government projects succeed, says cio. Computer Weekly 22, 4.

Damljanovic, D., Bontcheva, K., 2009. Towards enhanced usability of natural language interfaces to knowledge bases. Web 2.0 & Semantic Web, 105–133.

Damljanovic, D., Tablan, V., Bontcheva, K., 2008. A text-based query interface to owl ontologies. In: 6th Language Resources and Evaluation Conference (LREC), Marrakech, Morocco.

Davis, R., Shrobe, H., Szolovits, P., 1993. What is a knowledge representation? AI magazine 14 (1), 17.

Dias, G. P., June 2011. Local e-government information and service delivery, a survey of municipal websites in portugal. In: Rocha, A., Gonçalves, R., Cota, M. P., Reis, L. P. (Eds.), Sistemas e Tecnologias de Informação - Actas da 6^a Conferência Ibérica de Sistemas e Tecnologias de Informação. AISTI and UTAD, Chaves, Portugal, pp. 98–103.

Dias, G. P., Moreira, J. M., August 2008. Transparency corruption and ict (illustrated with portuguese cases). In: Vaccaro, A., Horta, H., Madsen, P. (Eds.), Transparency, Information and Communication Technology: Social Responsability and Accountability in Business and Education. Philosophy Documentation Center, Charlottesville, Virginia, USA, pp. 151–160.

Dunning, T., 1994. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19 (1), 61–74.

Embley, D. W., 2004. Toward semantic understanding – an approach based on information extraction ontologies. In: Proceedings of the 15th Australasian database conference-Volume 27. Australian Computer Society, pp. 3–12.

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., Yates, A., May 2004. Web-scale information extraction in knowitall (preliminary results). In: Feldman, S. I., Uretsky, M., Najork, M., Wills, C. E. (Eds.), WWW '04 - Proceedings of the 13th International World Wide Web Conference. Association for Computational Linguistics, New York, NY, USA, pp. 100–110.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., Yates, A., 2005. Unsupervised named-entity extraction from the web: An experimental study. Artificial Intelligence 165 (1), 91–134.

European Commission, December 2010. The european egovernment action plan 2011-2015: Harnessing ICT to promote smart, sustainable & innovative government. in Internet, accessed February 2011.
URL http://ec.europa.eu/information_society/activities/egovernment/action_plan_2011_2015/docs/action_plan_en_act_part1_v2.pdf

European Commission, February 2011. ICT for Government and Public Services | Europa - Information Society. in Internet, accessed February 2011.
URL http://ec.europa.eu/information_society/activities/egovernment/index_en.htm

Florini, A. M., April 1999. Does the invisible hand need a transparent glove? the politics of transparency. In: 11th Annual World Bank Conference on Development Economics. Carnegie Endowment for International Peace, Washington, USA.

Fonou-Dombeu, J. V., Huisman, M., 2011. Combining ontology development methodologies and semantic web platforms for e-government domain ontology development. International Journal of Web & Semantic Technology (IJWesT) 2 (2), 12–25.

Fountain, J. E., 2001. Public sector: Early stage of a deep transformation. In: Litan, R. E., Rivlin, A. M. (Eds.), The Economic Payoff from the Internet Revolution. Brookings Institution Press, Washington, DC, USA.

Francis, W. N., Kucera, H., 1979. Brown corpus manual. Letters to the Editor 5 (2), 7.

Freitas, C., Afonso, S., 2008. Bíblia florestal: Um manual linguístico da floresta sintá(c)tica. in Internet, accessed April 2013.
URL http://www.linguateca.pt/Floresta/BibliaFlorestal/

Freitas, C., Rocha, P., Bick, E., September 2008. Floresta sintá(c)tica: Bigger, thicker and easier. In: Teixeira, A., de Lima, V., de Oliveira, L., Quaresma, P. (Eds.), PROPOR 2008 - Proceedings of the International Conference on Computational Processing of the Portuguese

Language. Vol. 5190 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 216–219.

Frissen, V., Millard, J., Huijboom, N., Iversen, J. S., Kool, L., Kotterink, B., van Lieshout, M., van Staden, M., van der Duin, P., September 2007. The future of egovernment: An exploration of ict-driven models of egovernment for the EU in 2020. Tech. rep., European Commission, Joint Research Centre, Institute for Prospective Technological Studies, Seville, Spain.

Gaizauskas, R., Wilks, Y., 1998. Information extraction: Beyond document retrieval. Journal of Documentation 54 (1), 70–105.

Gatius, M., Mangham, A., Lesmo, L., May 2006. The hops project: Developing transformational government services. In: Cunningham, P., Cunningham, M. (Eds.), IST-Africa Conference Proceedings. IIMC - International Information Management Corporation, Pretoria, South Africa.

Gillick, D., June 2009. Sentence boundary detection and the problem with the u.s. In: Proceedings of the NAACL HLT 2009: Short Papers. Association for Computational Linguistics, pp. 241–244.

Giménez, J., Màrquez, L., May 2004. SVMTool: A general pos tagger generator based on support vector machines. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal.

Gómez-Pérez, A., Ortiz-Rodríguez, F., Villazón-Terrazas, B., November 2005. Legal ontologies for the spanish e-government. In: Marín, R., Onaindia, E., Bugarín, A., Reyes, J. S. (Eds.), CAEPIA 2005 - Proceedings of the 11th Conference of the Spanish Association for Artificial Intelligence. U. Coruña, AEPIA and USC, Springer Berlin / Heidelberg, pp. 301–310.

Goodwin, J., Dolbear, C., Hart, G., 2008. Geographical linked data: The administrative geography of great britain on the semantic web. Transactions in GIS 12, 19–30.

Gorin, A. L., Riccardi, G., Wright, J. H., 1997. How may i help you? Speech Communication 23 (1-2), 113–127.

Grefenstette, G., Tapanainen, P., 1994. What is a word, what is a sentence? problems of tokenization. In: Kiefer, F., Kiss, G., Pajzs, J. (Eds.), COMPLEX '94 - Proceedings of the 3rd International Conference on Computational Lexicography. Budapest, Hungary, pp. 79–87.

Gruber, T. R., June 1993. A translation approach to portable ontology specifications. Knowledge Acquisition 5 (2), 199–220.

Guarino, N., June 1998. Formal ontology and information systems. In: FOIS 98 - Proceedings of the International Conference on Formal Ontology in Information Systems. IOS Press, Amsterdam, pp. 3–15.

Güngör, T., 2010. Part-of-speech tagging. In: Indurkhya, N., Damerau, F. J. (Eds.), Handbook of Natural Language Processing, Second Edition. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA.

Haase, P., Lewen, H., Studer, R., Tran, D. T., Erdmann, M., d'Aquin, M., Motta, E., 2008. The neon ontology engineering toolkit. In: Korn, J. (Ed.), WWW 2008 Developers Track – 17th International World Wide Web Conference.

Hall, J., Nilsson, J., Nivre, J., Eryiğit, G., Megyesi, B., Nilsson, M., Saers, M., June 2007. Single malt or blended? a study in multilingual parser optimization. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007. Association for Computational Linguistics, Prague, Czech Republic, pp. 933–939.

Handschuh, S., Staab, S., Studer, R., 2003. Leveraging metadata creation for the semantic web with cream. Ki 2003: Advances in Artificial Intelligence, 19–33.

Hardy, C. A., Williams, S. P., 2011. Assembling e-government research designs: A transdisciplinary view and interactive approach. Public Administration Review 71 (3), 405–413.

Harman, D., Liberman, M., 1993. Tipster complete. Corpus number LDC93T3A, Linguistic Data Consortium, Philadelphia.

Heeks, R., 2003. Most e-government-for-development projects fail: how can risks be reduced? Institute for Development Policy and Management, University of Manchester Manchester.

Heeks, R., 2008. ICT4D 2.0: The next phase of applying ICT for international development. Computer 41 (6), 26–33.

Heflin, J., Hendler, J., 2001. A portrait of the semantic web in action. Intelligent Systems, IEEE 16 (2), 54–59.

Heiler, S., June 1995. Semantic interoperability. ACM Computing Surveys 27 (2), 271–273.

Hendriks, F., 2003. Public, administration, theory and learning: Interaction research as interpretation. Administrative Theory & Praxis 25 (3), 393–408.

Ho, S.-C., Kauffman, R. J., Liang, T.-P., 2011. Internet-based selling technology and e-commerce growth: a hybrid growth theory approach with cross-model inference. Information Technology and Management 12 (4), 409–429.

Hobson, S. F., Anand, R., Yang, J., Lee, J., 2011. Towards interoperability in municipal government: a study of information sharing practices. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (Eds.), Human-Computer Interaction - INTERACT 2011, Proceedings of the 13th IFIP TC 13 International Conference. Vol. 6946 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Lisbon, Portugal, pp. 233–247.

Hogben, G., 2007. Security issues and recommendations for online social networks. ENISA position paper.

Hornbæk, K., 2006. Current practice in measuring usability: Challenges to usability studies and research. International journal of human-computer studies 64 (2), 79–102.

IEEE, March 2010. Standards glossary. in Internet, accessed August 2011.
URL http://www.ieee.org/education_careers/education/standards/standards_glossary.html

INE, June 2012. Censos 2011 - resultados provisórios. in Internet, accessed June 2012.
URL http://www.ine.pt/scripts/flex_provisorios/Main.html

Iriberri, A., Leroy, G., August 2007. Natural language processing and e-government: Extracting reusable crime report information. In: IRI - Proceedings of the IEEE International Conference on Information Reuse and Integration. IEEE, IEEE Systems, Man, and Cybernetics Society, Las Vegas, USA, pp. 221–226.

ISO 9241-11, 1998. Ergonomic requirements for office work with visual display terminals (vdts). The international organization for standardization.

Jeff, B., 2000. At the dawn of e-government: the citizen as customer. Government Finance Review 16 (5), 15.

Jurafsky, D., Martin, J. H., January 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd Edition. Prentice Hall, New York, NY, USA.

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E., Swick, R. R., 2002. Annotea: an open rdf infrastructure for shared web annotations. Computer Networks 39 (5), 589–608.

Kalyanpur, A., Parsia, B., Sirin, E., Grau, B. C., Hendler, J., 2006. Swoop: A web ontology editing browser. Web Semantics: Science, Services and Agents on the World Wide Web 4 (2), 144–153.

Kaufmann, E., Bernstein, A., 2007. How useful are natural language interfaces to the semantic web for casual end-users? In: Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007. Springer, pp. 281–294.

Kaufmann, E., Bernstein, A., Fischer, L., 2007. Nlp-reduce: A "naive" but domain-independent natural language interface for querying ontologies. ESWC.

Kaufmann, E., Bernstein, A., Zumstein, R., 2006. Querix: A natural language interface to query ontologies based on clarification dialogs. In: 5th International Semantic Web Conference (ISWC 2006). pp. 980–981.

Kiss, T., Strunk, J., 2006. Unsupervised multilingual sentence boundary detection. Computational Linguistics 32 (4), 485–525.

Klinov, P., June 2008. Pronto: A non-monotonic probabilistic description logic reasoner. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (Eds.), The Semantic Web: Research and Applications - Proceedings of the 5th European Semantic Web Conference. Vol. 5021 of Lecture Notes in Computer Science. Springer-Verlag Berlin / Heidelberg, pp. 822–826.

Knublauch, H., Fergerson, R., Noy, N., Musen, M., November 2004. The protégé owl plugin: An open development environment for semantic web applications. In: McIlraith, S., Plexousakis, D., van Harmelen, F. (Eds.), The Semantic Web - ISWC 2004 - Proceedings of the 3rd International Semantic Web Conference. Vol. 3298 of Lecture Notes in Computer Science. Springer-Verlag Berlin / Heidelberg, pp. 229–243.

Kübler, S., McDonald, R., Nivre, J., 2009. Dependency parsing - Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Kuhn, T., 2013. A survey and classification of controlled natural languages. Computational Linguistics.

Lee, L., 2004. "I'm sorry Dave, I'm afraid I can't do that": Linguistics, statistics, and natural language processing circa 2001. In: Committee on the Fundamentals of Computer Science: Challenges and Computer Science Opportunities and National Research Council Telecommunications Board (Ed.), Computer Science: Reflections on the Field, Reflections from the Field. The National Academies Press, Washington DC, USA, pp. 111–118.

Leighninger, M., 2011. Citizenship and governance in a wild, wired world: How should citizens and public managers use online tools to improve democracy? National Civic Review 100 (2), 20–29.

Lewis, D., 1970. General semantics. Synthese 22 (1-2), 18–67.

Lewis, J., 1995. Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction 7 (1), 57–78.

Li, Y., Yang, H., Jagadish, H. V., July 2005. Nalix: an interactive natural language interface for querying xml. In: Özcan, F. (Ed.), SIGMOD Conference. Association for Computational Linguistics, New York, NY, USA, pp. 900–902.

Lopez, V., Uren, V., Motta, E., Pasin, M., 2007. Aqualog: An ontology-driven question answering system for organizational semantic intranets. Web Semantics: Science, Services and Agents on the World Wide Web 5 (2), 72–105.

Lopez-Pellicer, F. J., Chaves, M., Rodrigues, C., Silva, M. J., 2009. Geographic ontologies production in grease-ii. Tech. Rep. 09-18, Universidade de Lisboa, Faculdade de Ciências, LASIGE.

Lörincz, B., Tinholt, D., van der Linden, N., Colclough, G., Cave, J., Schindler, R., Cattaneo, G., Lifonti, R., Jacquet, L., Millard, J., December 2010. Digitizing public services in europe: Putting ambition into action - 9th benchmark measurement. Tech. rep., Prepared by: Capgemini, RAND Europe, IDC, SOGETI and DTi. For: European Commission Directorate General for Information Society and Media.

Maedche, A., Neumann, G., Staab, S., 2003. Bootstrapping an ontology-based information extraction system. Studies In Fuzziness And Soft Computing 111, 345–362.

Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., 1999. Performance measures for information extraction. In: Proceedings of DARPA Broadcast News Workshop. pp. 249–252.

Malmö Declaration, November 2009. Ministerial declaration on egovernment. Accessed February 2011.
URL http://www.epractice.eu/files/MalmoMinisterialDeclaration2009.pdf

Manning, C. D., February 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In: Gelbukh, A. (Ed.), Computational Linguistics and Intelligent Text Processing - 12th International Conference CICLing. Vol. 6608 of Lecture Notes in Computer Science. Springer, pp. 171–189.

Marcus, M. P., Marcinkiewicz, M. A., Santorini, B., June 1993. Building a large annotated corpus of english: The penn treebank. Computational Linguistics 19 (2), 313–330.

Màrquez, L., Carreras, X., Litkowski, K. C., Stevenson, S., December 2008. Semantic role labeling: An introduction to the special issue. Computational Linguistics 34 (2), 145–159.

Màrquez, L., Klein, D. (Eds.), 2006. CoNLL-X - Proceedings of the Tenth Conference on Computational Natural Language Learning. Omnipress Inc.

Marrafa, P., Amaro, R., Mendes, S., 2011. Wordnet. pt global: extending wordnet. pt to portuguese varieties. In: Proceedings of the First Workshop on Algorithms and Resources

for Modelling of Dialects and Language Varieties. Association for Computational Linguistics, pp. 70–74.

McDowell, L., Etzioni, O., Gribble, S., Halevy, A., Levy, H., Pentney, W., Verma, D., Vlasseva, S., 2003. Mangrove: Enticing ordinary people onto the semantic web via instant gratification. In: The Semantic Web-ISWC 2003. Vol. 2870 of Lecture Notes in Computer Science. Springer, pp. 754–770.

McDowell, L. K., Cafarella, M., November 2006. Ontology-driven information extraction with ontosyphon. In: Cruz, I. F., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L. (Eds.), The Semantic Web - ISWC2006. Vol. 4273 of Lecture Notes in Computer Science. Springer, pp. 428–444.

McNaught, J., Black, W., 2006. Information extraction. In: Ananiadou, S., McNaught, J. (Eds.), Text Mining for Biology and Biomedicine. Artech House.

McQueen, J., Vasques, C., Brakebill, J., Roberts, D., Parston, G., January 2009. Leadership in customer service: Creating shared responsibility for better outcomes. Tech. rep., Accenture.

Mikheev, A., 2002. Periods, capitalized words, etc. Computational Linguistics 28 (3), 289–318.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J., 1990. Introduction to wordnet: An on-line lexical database*. International journal of lexicography 3 (4), 235–244.

Misuraca, G. C., 2009. e-government 2015: exploring m-government scenarios, between ict-driven experiments and citizen-centric implications. Technology Analysis & Strategic Management 21 (3), 407–424.

Mohamed, F. Z. M., Muthaiyah, S. S., Nassirtoussi, A. K., 2011. Tm: a development technique for e-government 2.0 portals. WSEAS Transactions on Information Science and Applications 8 (2), 96–108.

Mota, C., Santos, D. (Eds.), December 2008. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca.

Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Linguisticae Investigationes 30, 3–26.

Nass, C., Brave, S., 2005. Wired for speech: How voice activates and advances the human-computer relationship. MIT press.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D., 2007. The CoNLL 2007 shared task on dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007. pp. 915–932.

Nivre, J., Hall, J., Nilsson, J., Eryiğit, G., Marinov, S., June 2006. Labeled pseudo-projective dependency parsing with support vector machines. In: CoNLL-X - Proceedings of the 10th Conference on Computational Natural Language Learning. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 221–225.

Noy, N., Fergerson, R., Musen, M., October 2000. The knowledge model of protégé-2000: Combining interoperability and flexibility. In: Dieng, R., Corby, O. (Eds.), EKAW 2000 - Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management. Vol. 1937 of Lecture Notes in Computer Science. INRIA, AFIA, MLnet in collaboration with AAAI, Springer-Verlag Berlin / Heidelberg, pp. 69–82.

OECD, 2005. Modernising government: the way forward. Organisation for Economic Co-operation and Development.

Ogden, W. C., Bernick, P., 1997. Using natural language interfaces. Handbook of human-computer interaction, 137–161.

Oliveira, H. G., Santos, D., Gomes, P., 2009. Relations extracted from a portuguese dictionary: results and first evaluation. In: Local Proc. 14th Portuguese Conf. on Artificial Intelligence (EPIA). pp. 541–552.

Oliveira, H. G., Santos, D., Gomes, P., Seco, N., 2008. Papel: a dictionary-based lexical ontology for portuguese. In: Computational Processing of the Portuguese Language. Springer, pp. 31–40.

Oracle, 2013. Java servlet technology. Accessed January 2013.
URL http://www.oracle.com/technetwork/java/index-jsp-135475.html

Oren, E., Möller, K., Scerri, S., Handschuh, S., Sintek, M., 2006. What are semantic annotations. Tech. rep., Digital Enterprise Research Institute at National University of Ireland, German Research Center for Artificial Intelligence (DFKI).

Ortiz-Rodríguez, F., Palma, R., Villazón-Terrazas, B., September 2007. Egoir: ontology-based information retrieval intended for egovernment. In: Koschke, R., Herzog, O., Rödiger, K.-H., Ronthaler, M. (Eds.), Informatik 2007 - Proceedings of the Annual Meeting of the Society for computer science. Springer, Bremen, Germany, pp. 237–241.

Osimo, D., 2008. Web 2.0 in government: Why and how? Tech. Rep. EUR 23358, Institute for Prospectice Technological Studies (IPTS), Joint Research Centre (JRC), European Commission.

Padró, L., Collado, M., Reese, S., Marina Lloberes and, I. C., May 2010. Freeling 2.1: Five years of open-source language processing tools. In: Calzolari, N., Choukri, K., Maegaard,

B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta.

Paiva Dias, G., Rafael, J. A., January 2007. A simple model and a distributed architecture for realizing one-stop e-government. Electronic Commerce Research and Applications 6 (1), 81–90.

Palmer, D. D., Hearst, M. A., 1997. Adaptive multilingual sentence boundary disambiguation. Computational Linguistics 23 (2), 241–267.

Paul, D. B., Baker, J. M., 1992. The design for the wall street journal-based csr corpus. In: Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, pp. 357–362.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M., 2003. Kim – semantic annotation platform. In: Proceedings of the Semantic Web-ISWC 2003. Springer, pp. 834–849.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A., 9 2004. Kim - a semantic platform for information extraction and retrieval. Natural Language Engineering 10, 375–392.

Quinlan, J. R., 1983. Learning efficient classification procedures and their application to chess end games. In: Michalski, R. S., Carbonell, J. G., Mitchel, T. M. (Eds.), Machine learning: An artificial intelligence approach. Vol. 1. Tioga Publishing, Palo Alto, CA, pp. 463–482.

Ranchhod, E. M., 2005. Using corpora to increase portuguese mwe dictionaries. tagging mwe in a portuguese corpus. In: Proceedings from The Corpus Linguistics Conference Series. Vol. 1.

Ratnaparkhi, A., May 1996. A maximum entropy model for part-of-speech tagging. In: Brill, E., Church, K. (Eds.), Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 133–142.

Reddick, C. G., 2005. Citizen interaction with e-government: From the streets to servers? Government Information Quarterly 22 (1), 38–57.

Relyea, H. C., 2002. E-gov: Introduction and overview. Government Information Quarterly 19 (1), 9–36.

Reynar, J. C., Ratnaparkhi, A., 1997. A maximum entropy approach to identifying sentence boundaries. In: Proceedings of the fifth conference on Applied natural language processing. ANLC '97. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 16–19.

Richardson, S. D., Dolan, W. B., Vanderwende, L., 1998. Mindnet: acquiring and structuring semantic information from text. In: Proceedings of the 17th international conference on Computational linguistics-Volume 2. Association for Computational Linguistics, pp. 1098–1102.

Riley, M. D., 1989. Some applications of tree-based modelling to speech and language. In: Proceedings of the workshop on Speech and Natural Language. HLT '89. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 339–352.

Riloff, E., 1999. Information extraction as a stepping stone toward story understanding. In: Understanding Language: Understanding Computational Models of Reading. MIT Press, pp. 435–460.

Rodrigues, M., Dias, G. P., Teixeira, A., June 2010a. Automatic extraction and representation of geographic entities in egovernment. In: Rocha, A., Sexto, C. F., Reis, L. P., Cota, M. P. (Eds.), Sistemas y Tecnologías de Informacíon - Actas de la 5a Conferencia Ibérica de Sistemas y Tecnologías de Informacíon. AISTI, GIS-T and USC, Santiago de Compostela, Spain, pp. 160–163.

Rodrigues, M., Dias, G. P., Teixeira, A., April 2010b. Human language technologies for e-gov. In: Filipe, J., Cordeiro, J. (Eds.), WEBIST 2010, Proceedings of the 6th International Conference on Web Information Systems and Technologies. Organized by INSTICC in co-operation with WfMC and ACM SIGMIS, Valencia, Spain, pp. 400–403.

Rodrigues, M., Dias, G. P., Teixeira, A., November 2010c. Knowledge extraction from minutes of portuguese municipalities meetings. In: Mateo, C. G., Diaz, F. C., Pazó, F. M. (Eds.), FALA 2010: VI Jornadas en Tecnologia del Habla and II Iberial SLTech - Speech and Language Technologies for Iberian Languages. MTG, RTTH and ISCA SIG-IL, Vigo, Spain, pp. 51–54.

Rodrigues, M., Dias, G. P., Teixeira, A., 2011a. Criação e Acesso a Informação Semântica Aplicada ao Governo Eletrónico. Linguamática 3 (2), 55–68.

Rodrigues, M., Dias, G. P., Teixeira, A., October 2011b. Ontology Driven Knowledge Extraction System with Application in e-Government. In: Proc. of the 15th Portuguese Conference on Artificial Intelligence. Lisboa, Portugal, pp. 760–774.

Rodrigues, M., Dias, G. P., Teixeira, A., 2013. Towards e-government information platforms for enterprise 2.0. In: Cruz-Cunha, M. M., Moreira, F., ao Varajão, J. (Eds.), Handbook of Research on Enterprise 2.0: Technological, Social, and Organizational Dimension. IGI Global.

Rohrer, C., 2008. Desirability studies. Accessed May 2012.
URL http://www.xdstrategy/com/2008/10/28/desirability-studies

Sag, I. A., Wasow, T., 1999. Syntactic theory: A formal introduction. Vol. 92. Center for the Study of Language and Information.

Saggion, H., Funk, A., Maynard, D., Bontcheva, K., 2007. Ontology-based information extraction for business intelligence. In: The Semantic Web. Vol. 4825 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 843–856.

Sanchez, D., Isern, D., Millan, M., 2011. Content annotation for the semantic web: an automatic web-based approach. Knowledge and Information Systems 27 (3), 393–418.

Santos, L. D., Amaral, L., 2008. Presença na internet das câmaras municipais portuguesas em 2007: estudo sobre local egovernment em portugal. Tech. rep., Gávea - Laboratório de Estudo e Desenvolvimento da Sociedade da Informação do Departamento de Sistemas de Informação da Universidade do Minho.

Schmid, H., September 1994. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing. Manchester, UK.

Scholl, H. J., Klischewski, R., 2007. E-government integration and interoperability: framing the research agenda. International Journal of Public Administration 30 (8-9), 889–920.

Schroeter, R., Hunter, J., Kosovic, D., 2003. Vannotea: A collaborative video indexing, annotation and discussion system for broadband networks. In: Knowledge capture. ACM Press (Association for Computing Machinery), pp. 1–8.

Schwitter, R., 2002. English as a formal specification language. In: Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on. IEEE, pp. 228–232.

Seo, H.-J., Lee, Y. S., Oh, J. H., 2009. Does ICT investment widen the growth gap? Telecommunications Policy 33 (8), 422–431.

Silla, C. N., Kaestner, C. A. A., February 2004. An analysis of sentence boundary detection systems for english and portuguese documents. In: Gelbukh, A. F. (Ed.), CICLing - Proceedings of the 5th International Conference on Computational Linguistics and Intelligent Text Processing. Vol. 2945 of Lecture Notes in Computer Science. Springer, pp. 135–141.

Sirin, E., Parsia, B., June 2004. Pellet: An owl dl reasoner. In: Haarslev, V., Möller, R. (Eds.), DL 2004 - Proceedings of the 2004 International Workshop on Description Logics. Vol. 104 of CEUR Workshop Proceedings. pp. 212–213.

Sowa, J. F., 2000. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA, USA.

Sroga, M., October 2008. Access-egov - personal assistant of public services. In: Ganzha, M., Paprzycki, M., Pełech-Pilichowski, T. (Eds.), IMCSIT - Proceedings of the International Multiconference on Computer Science and Information Technology. PTI - Polish Information Processing Society, IEEE Computer Society Press, Los Alamitos, CA, USA, pp. 421–427.

Suchanek, F. M., Ifrim, G., Weikum, G., July 2006. Leila: Learning to extract information by linguistic analysis. In: Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Association for Computational Linguistics, Sydney, Australia, pp. 18–25.

Suchanek, F. M., Kasneci, G., Weikum, G., May 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In: WWW '07 - Proceedings of the 16th International World Wide Web Conference. ACM, New York, NY, USA, pp. 697–706.

Todirascu, A., Romary, L., Bekhouche, D., 2002. Vulcain – an ontology-based information extraction system. In: Natural Language Processing and Information Systems. Vol. 2553 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 64–75.

Travis, D., 2008. Measuring satisfaction: Beyond the usability questionnaire. Accessed May 2012.
URL http://www.userfocus.co.uk/articles/satisfaction.html

Tullis, T., Stetson, J., 2004. A comparison of questionnaires for assessing website usability. In: Usability Professional Association Conference.

United Nations, 2010. United nations e-government survey 2010 - leveraging e-government at a time of financial and economic crisis. Tech. rep., United Nations Department of Economic and Social Affairs, New York, USA.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F., 2006. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semantics: science, services and agents on the World Wide Web 4 (1), 14–28.

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., Ciravegna, F., 2002. Mnm: Ontology driven semi-automatic and automatic support for semantic markup. Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web 2473, 213–221.

Volz, R., Handschuh, S., Staab, S., Stojanovic, L., Stojanovic, N., 2004. *Unveiling the hidden bride*: deep annotation for mapping and migrating legacy data to the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 1 (2), 187–206.

Wang, C., Xiong, M., Zhou, Q., Yu, Y., June 2007a. Panto: A portable natural language interface to ontologies. In: Franconi, E., Kifer, M., May, W. (Eds.), ESWC2007 - Proceedings of the 4th European Semantic Web Conference. Vol. 4519 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 473–487.

Wang, H., Song, Y., Hamilton, A., Curwell, S., 2007b. Urban information integration for advanced e-planning in europe. Government Information Quarterly 24 (4), 736–754.

Wang, P., Zheng, J. G., Fu, L., Patton, E. W., Lebo, T., Ding, L., Liu, Q., Luciano, J. S., McGuinness, D. L., 2011. A semantic portal for next generation monitoring systems. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (Eds.), The Semantic Web - ISWC 2011, Proceedings of the 10th International Semantic Web Conference. Vol. 7032 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Bonn, Germany, pp. 253–268.

Wang, X. H., Zhang, D. Q., Gu, T., Pung, H. K., March 2004. Ontology based context modeling and reasoning using owl. In: Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops. PERCOMW '04. IEEE Computer Society, Washington, DC, USA, pp. 18–22.

Weibel, S., Kunze, J., Lagoze, C., Wolf, M., 1998. Dublin core metadata for resource discovery. Internet Engineering Task Force RFC 2413.

Weingarten, F. W., 1994. Public interest and the NII. Communications of the ACM 37 (3), 17–19.

Whitelaw, C., Kehlenbeck, A., Petrovic, N., Ungar, L. H., October 2008. Web-scale named entity recognition. In: CIKM 2008 - Proceedings of the 17th ACM conference on Information and knowledge management. ACM, New York, NY, USA, pp. 123–132.

Williams, D., Kelly, G., Anderson, L., 2004. Msn 9: new user-centered desirability methods produce compelling visual design. In: CHI'04 extended abstracts on Human factors in computing systems. ACM, pp. 959–974.

Wimalasuriya, D. C., Dou, D., June 2010. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science 36 (3), 306–323.

World Bank, February 2011. e-government - definition of e-government. in Internet, accessed February 2011.
URL http://go.worldbank.org/M1JHE0Z280

Wu, F., Hoffmann, R., Weld, D. S., August 2008. Information extraction from wikipedia: Moving down the long tail. In: Li, Y., Liu, B., Sarawagi, S. (Eds.), KDD '08 - Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp. 731–739.

Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O., Soderland, S., April 2007. Textrunner: Open information extraction on the web. In: Sidner, C. L., Schultz, T., Stone, M., Zhai, C. (Eds.), NAACL-HLT (Demonstrations) - Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, USA, pp. 25–26.

Yildiz, B., Miksch, S., 2007. ontox - a method for ontology-driven information extraction. In: Computational Science and Its Applications – ICCSA 2007. Vol. 4707 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 660–673.

A

LABEL-LEX-sw Tag Set Conversion

Table A.1: Conversion of LABEL-LEX-sw tags to CoNLL-X (Bosque) tags.

LABEL-LEX-sw POS tags	CoNLL-X POS tags (CPOSTAG:POSTAG)
A_NC	adj:adj, prop:prop, n:n
ADJ	adj:adj
ADV, PREPXADV	adv:adv
X, XC	adv:adv, n:n, prp:prp
CONJ	conj:conj-c, conj:conj-s
DET, DETXDET	art:art, num:num
DETXPRO, PREPXPRO, PREPXPROXPRO, PRO, PROXPRO	pron:pron-pers, pron:pron-indp, pron:pron-det
PREPXDET, Vmf	pron:pron-pers
PREP, PREPXPREP	prp:prp
IN, INTERJ, ONOM	in:in
N, NC, PFX	n:n
NP, NPM	prop:prop
CP, VC, VF, VFC, VI, VIN, Vmc, VP, VPC, VPI, VPIC, VPMP, VPP	v:v-fin
VG	v:v-ger
VII, VIP	v:v-inf
VPMPCI, VPPA	v:v-pcp
<no_tag>, PUNCT1A, PUNCT1B, PUNCT1C, PUNCT2D, PUNCT2E, PUNCT2F, PUNCTG, PUNCTH, PUNCTMR	punc:punc

B

Semantic Model Creation Algorithm

Algorithm 1 Training procedure

```
//Finding examples in sentence graphs
for all g ∈ annotatedSentenceSet do
  for all e ∈ exampleSet do
    if e ⊂ s then
      relationModel(e.relation).addExample(s.graph, e.subject.text, e.object.text)
      entity-ofModel(e.subject.class).add(s.graph, e.subject.text)
      entity-ofModel(e.object.class).add(s.graph, e.object.text)
    end if
  end for
end for

//Information share. Optional step
for all m ∈ allModelSet do
  disjointModelSet = m.getModelsOfDisjointClasses()
  for all dm ∈ disjointModelSet do
    dm.addCounterExamples(m.getExamples())
  end for
  superModelSet = m.getModelsOfSuperClasses()
  for all sm ∈ superModelSet do
    sm.addExamples(m.getExamples())
    sm.addCounterExamples(m.getCounterExamples())
  end for
end for

//Training entity-of models
for all eom ∈ entity-ofModelSet do
  eom.train()
end for

//Training relation models
for all rel ∈ relationModelSet do
  repeat
    rel.train()
    count ← 0
    for all s ∈ documentSentences do
      for all p ∈ s.allPossibleWordPairs() do
        if rel.classify(p) > threshold and p ∉ exampleSetOfRelation(rel) then
          rel.addCounterExample(s.graph, p.subject.text,p.object.text)
          count++
        end if
      end for
    end for
  until count < countThreshold
end for
```

C

Application Setup

The application setup described here assumes that the target application is about making available information written in Portuguese. The tasks required to prepare the proof-of-concept prototype for a given application are:

1. Define the ontology - the ontology can be completely defined or can be built by reusing existing ontologies. Both cases are supported by Protégé, being the only requirement having the ontology defined using OWL. It is advisable to have a semantic reasoner evaluating the ontology before its usage.

2. Select documents for providing examples of concepts - these documents will be used to provide semantic relation examples. It is important to select a representative set of documents. By representative set of documents is meant a set of documents that contains most, is not all, information types that the system will be required to process, and in a similar proportion that is expected during system usage.

3. Provide relevant semantic relation examples - example annotation should be done by more than one person. This implies defining precisely what should be annotated to prevent different styles of annotations. For instance, let us consider that are wanted examples about exemptions requested and/or granted. The text about the rules for granting exemptions is not relevant, "... citizens need to present the ID card to request an exemption ..."; but a sentence about an exemption outcome is "... exemption was granted to citizen with ID card 123456 ...".

4. Create semantic extraction models - having annotated examples of semantic relations and the documents where they were found, is now possible to create semantic extraction models. The semantic extraction models and the ontology are the outputs of task preparation as, from now on, the prototype behavior depends on these.

In runtime, updating the knowledge base with new documents starts with processing those documents using the NLP pipeline. Then, the semantic extraction models acts upon the NLP output obtaining triples reflecting the documents information. It is advisable to check if information is being correctly extracted. If not, repeat prototype preparation steps 2 to 4, preferably including as sample documents those where the semantic extraction failed more. When convenient, the system can be stopped in order to feed the knowledge base with new triples reflecting the information extracted from the new set of documents.

D

Product Reaction Cards Translation

Table D.1: Translation of product reaction cards.

Original (en)	Translation (pt)	Original (en)	Translation (pt)
Approachable	Abordável	Impersonal	Impessoal
Accessible	Acessível	Impressive	Impressionante
Advanced	Avançado	Incomprehensible	Incompreensível
Annoying	Irritante	Inconsistent	Inconsistente
Appealing	Apelativo	Ineffective	Ineficaz
Attractive	Atraente	Innovative	Inovador
Boring	Maçador	Inspiring	Inspirador
Business-like	Empresarial	Integrated	Integrado
Busy	Ocupante	Intimidating	Intimidante
Calm	Calmo	Intuitive	Intuitivo
Clean	Limpo	Inviting	Convidativo
Clear	Claro	Irrelevant	Irrelevante
Collaborative	Colaborativo	Low maintenance	Baixa manutenção
Comfortable	Confortável	Meaningful	Significativo
Compatible	Compatível	Motivating	Motivador
Compelling	Compelativo	Not secure	Não seguro
Complex	Complexo	Not valuable	Sem valor
Comprehensive	Abrangente	Novel	Novo
Confident	Confiante	Old	Velho

Continues on next page

Table D.1 – continued from previous page

Original (en)	Translation (pt)	Original (en)	Translation (pt)
Confusing	Confuso	Optimistic	Otimista
Connected	Ligado	Ordinary	Vulgar
Consistent	Consistente	Organized	Organizado
Controllable	Controlável	Overbearing	Arrogante
Convenient	Conveniente	Overwhelming	Esmagador
Creative	Criativo	Patronizing	Paternalista
Customizable	Personalizável	Personal	Pessoal
Cutting-edge	De ponta (tecnologia/sistema)	Poor quality	Má qualidade
Dated	Datado	Powerful	Poderoso
Desirable	Desejável	Predictable	Previsível
Difficult	Difícil	Professional	Profissional
Disconnected	Desligado	Relevant	Relevante
Disruptive	Disruptivo	Reliable	Fiável
Distracting	Distrator	Responsive	Responsivo
Dull	Aborrecido	Rigid	Rígido
Easy to use	Fácil de usar	Satisfying	Satisfaz
Effective	Eficaz	Secure	Seguro
Efficient	Eficiente	Simplistic	Simplista
Effortless	Sem esforço	Slow	Lento
Empowering	Empossante	Sophisticated	Sofisticado
Energetic	Enérgico	Stable	Estável
Engaging	Envolvente	Sterile	Estéril
Entertaining	Entretém	Stimulating	Estimulante
Enthusiastic	Entusiasta	Straightforward	Direto
Essential	Essencial	Stressful	Enervante
Exceptional	Excepcional	Time-consuming	Demorado
Exciting	Emocionante	Time-saving	Poupa tempo
Expected	Esperado	Too technical	Muito técnico
Familiar	Familiar	Trustworthy	Confiável
Fast	Rápido	Unapproachable	Inacessível
Flexible	Flexível	Unattractive	Pouco atraente
Fragile	Frágil	Uncontrollable	Incontrolável
Fresh	Fresco	Unconventional	Não convencional
Friendly	Amigável	Understandable	Compreensível
Frustrating	Frustrante	Undesirable	Indesejável
Fun	Divertido	Unpredictable	Imprevisível
Gets in the way	Atrapalha	Unrefined	Não refinado
Hard to use	Complicado de usar	Usable	Utilizável
Helpful	Ajudante	Useful	Útil
High quality	Alta qualidade	Valuable	Valioso